On the infinite-dimensional symmetry group of the Davey–Stewartson equations

B. Champagne[a] and P. Winternitz
Centre de recherches mathématiques, Université de Montréal, C.P. 6128, Succ. A, Montréal, Québec, Canada, H3C 3J7

(Received 19 February 1987; accepted for publication 5 August 1987)

The Lie algebra of the group of point transformations, leaving the Davey–Stewartson equations (DSE's) invariant, is obtained. The general element of this algebra depends on four arbitrary functions of time. The algebra is shown to have a loop structure, a property shared by the symmetry algebras of all known $(2+1)$-dimensional integrable nonlinear equations. Subalgebras of the symmetry algebra are classified and used to reduce the DSE's to various equations involving only two independent variables.

I. INTRODUCTION

The purpose of this paper is to apply the method of symmetry reduction to the Davey–Stewartson equations (DSE's).[1] To do this we first obtain the group of Lie point symmetries leaving the DSE's invariant. We show that this group is infinite dimensional, study its structure, and determine its low-dimensional subgroups. The different subgroups are then used to reduce the DSE's to a lower-dimensional system.

We recall that the DSE's describe the propagation of two-dimensional water waves moving under the force of gravity in water of finite depth. We shall write these equations in the form

$$\begin{aligned} &i\Psi_t + \Psi_{xx} + \epsilon_1 \Psi_{yy} = \epsilon_2 |\Psi|^2 \Psi + \Psi w\,, \\ &w_{xx} + \delta_1 w_{yy} = \delta_2 (|\Psi|^2)_{yy}\,, \end{aligned} \tag{1.1}$$

where $\Psi(x,y,t)$ and $w(x,y,t)$ are a complex and real function, respectively, and δ_1, δ_2, ϵ_1, and ϵ_2 are real constants, with $\epsilon_1 = \pm 1$, $\epsilon_2 = \pm 1$. The subscripts in (1.1) denote partial derivatives.

For purely one-dimensional propagation (along the x axis) we have $\Psi_y = 0$ and can consider solutions with $w=0$. The DSE's (1.1) then reduce to the nonlinear Schrödinger equation

$$i\Psi_t + \Psi_{xx} = \epsilon_2 |\Psi|^2 \Psi\,. \tag{1.2}$$

The DSE's thus have the same relation to the nonlinear Schrödinger equation as the Kadomtsev–Petviashvili equation[2] has to the Korteweg–de Vries one, they provide a two-dimensional generalization in which the basic direction of wave propagation remains a privileged one.

The DSE's belong to the rather limited class of equations in more than $1+1$ dimensions that are exactly integrable[3] by inverse scattering techniques and their generalizations.[4–6] In particular, the DSE's were shown to have soliton and multisoliton solutions.[3]

Some recent work has been devoted to the study of symmetry groups of integrable equations in more than two dimensions.[7–12] Thus the Kadomtsev–Petviashvili equation,[7] the modified Kadomtsev–Petviashvili equation,[8] the potential Kadomtsev–Petviashvili equation,[12] and the integrable three-wave problem[10,11] all have infinite-dimensional symmetry groups. The corresponding infinite-dimensional Lie algebras all have a specific loop algebra structure. They all have Virasoro-type subalgebras and can be embedded into simple classical loop algebras of the $A_n^{(1)}$ type.[13] On the other hand, some of the multidimensional equations of the Jimbo–Miwa series,[14] which are integrable in a conditional sense,[9] have been shown to have infinite-dimensional Lie symmetry algebras that are not loop algebras.[9]

In Sec. II we present the symmetry algebra of the DSE's and exhibit its loop algebra structure by relating it to the algebra sl(7,C). We also obtain the group transformations by integrating the vector fields forming the symmetry algebra. In Sec. III we classify the one- and two-dimensional subalgebras of the DS algebra into conjugacy classes under the action of the DS group. The one-dimensional subalgebras are used in Sec. IV to reduce the DS equations to various integrable systems in $1+1$ dimensions.

II. THE SYMMETRY GROUP OF THE DAVEY–STEWARTSON EQUATIONS

A. The DS symmetry algebra

Standard procedures exist for determining the symmetry algebra of a system of differential equations.[15] They are so algorithmic that they have been successfully programmed using REDUCE,[16] MACSYMA,[8] or other symbolic languages. In order to be able to apply a previously written program,[8] we rewrite the DSE's (1.1) in a real form, setting $\psi = u + iv$. We obtain

$$\begin{aligned} &\Delta_1 \equiv u_t + v_{xx} + \epsilon_1 v_{yy} - \epsilon_2 v(u^2+v^2) - vw = 0, \\ &\Delta_2 \equiv -v_t + u_{xx} + \epsilon_1 u_{yy} - \epsilon_2 u(u^2+v^2) - uw = 0\,, \\ &\Delta_3 \equiv w_{xx} + \delta_1 w_{yy} \\ &\qquad - 2\delta_2[u u_{yy} + (u_y)^2 + v v_{yy} + (v_y)^2] = 0\,. \end{aligned} \tag{2.1}$$

An element of the symmetry algebra of (2.1) is written as

$$V = \eta_1\,\partial_x + \eta_2\,\partial_y + \eta_3\,\partial_t + \phi_1\,\partial_u + \phi_2\,\partial_v + \phi_3\,\partial_w\,, \tag{2.2}$$

[a] Present address: Department of Electrical Engineering, University of Toronto, Toronto, Ontario, Canada, M5S 1A4.

where η_i and ϕ_i ($i=1,2,3$) are functions of x, y, t, u, v, and w. These functions are obtained by solving the determining equations, that in turn follow from the equations

$$\mathrm{pr}^{(2)}V\cdot\Delta_j(x,y,t,u,v,w)|_{\Delta_i=0}=0,\quad i=1,2,3\,, \tag{2.3}$$

where $\mathrm{pr}^{(2)}V$ is the second prolongation[15] of the vector field V. Applying the program[8] we obtain the determining equations, a relatively simple system of linear partial differential equations for η_i and ϕ_i. By solving them we find that a general element of the symmetry algebra of the DSE's (2.1) can be written as

$$V=X(f)+Y(g)+Z(h)+W(m)\,, \tag{2.4}$$

where

$$\begin{aligned} X(f) &= f(t)\partial_t+[f'(t)/2](x\,\partial_x+y\,\partial_y-u\,\partial_u \\ &\qquad -v\,\partial_v-2w\,\partial_w)-[(x^2+\epsilon_1 y^2)/8] \\ &\qquad \times[f''(t)(v\,\partial_u-u\,\partial_v)+f'''(t)\partial_w]\,, \\ Y(g) &= g(t)\partial_x-[x/2] \\ &\qquad \times[g'(t)(v\,\partial_u-u\,\partial_v)+g''(t)\partial_w]\,, \\ Z(h) &= h(t)\partial_y-[\epsilon_1 y/2] \\ &\qquad \times[h'(t)(v\,\partial_u-u\,\partial_v)+h''(t)\partial_w]\,, \\ W(m) &= m(t)(v\,\partial_u-u\,\partial_v)+m'(t)\partial_w\,. \end{aligned} \tag{2.5}$$

The functions $g(t)$, $h(t)$, and $m(t)$ are arbitrary real-valued functions of class C^∞ over some time interval $T\subseteq\mathbf{R}$. The function $f(t)$ satisfies

$$f(t)=\begin{cases}\text{arbitrary}, & \text{if } \delta_1=-\epsilon_1=\pm1,\\ a+bt+ct^2, & \text{if } \delta_1\neq-\epsilon_1\end{cases} \tag{2.6a}$$

(a, b, and c are arbitrary real constants). The primes in (2.5) denote derivatives with respect to time t. The DSE's have been shown to be integrable precisely in the case when we have

$$\delta_1=-\epsilon_1\,, \tag{2.6b}$$

i.e., when $f(t)$ is allowed to be arbitrary. We shall mainly concentrate on this case. The commutation relations for the DS algebra (2.4), (2.5) are easy to obtain, namely

$$\begin{aligned} &[X(f_1),X(f_2)]=X(f_1f_2'-f_1'f_2)\,, \\ &[X(f),Y(g)]=Y(fg'-f'g/2)\,, \\ &[X(f),Z(h)]=Z(fh'-f'h/2)\,, \\ &[X(f),W(m)]=W(fm')\,, \\ &[Y(g_1),Y(g_2)]=-W(g_1g_2'-g_1'g_2)/2\,, \\ &[Z(h_1),Z(h_2)]=-\epsilon_1W(h_1h_2'-h_1'h_2)/2, \\ &[Y(g),Z(h)]=[Y(g),W(m)]=[Z(h),W(m)] \\ &\qquad =[W(m_1),W(m_2)]=0\,. \end{aligned} \tag{2.7}$$

We see that the DS Lie algebra L allows a Levi decomposition[17]

$$L=S\,\dot{+}\,N, \tag{2.8}$$

where $S=\{X(f)\}$ is a simple infinite-dimensional Lie algebra and $N=\{Y(g), Z(h), W(m)\}$ is the radical of L, which in this case happens to be a nilpotent ideal.

The "obvious" physical symmetries of the DSE's are obtained by restricting all the functions f, g, h, and m to be first-order polynomials. We then have

$$\begin{aligned} &P_0=X(1)=\partial_t,\quad P_1=Y(1)=\partial_x, \\ &P_2=Z(1)=\partial_y,\quad R_0=W(1)=v\,\partial_u-u\,\partial_v, \\ &D=X(t)=t\,\partial_t+(x\,\partial_x+y\,\partial_y \\ &\qquad -u\,\partial_u-v\,\partial_v)/2-w\,\partial_w\,, \\ &B_1=Y(t)=t\,\partial_x-x(v\,\partial_u-u\,\partial_v)/2, \\ &B_2=Z(t)=t\,\partial_y-\epsilon_1 y(v\,\partial_u-u\,\partial_v)/2\,, \\ &R_1=W(t)=t(v\,\partial_u-u\,\partial_v)+R_0\,. \end{aligned} \tag{2.9}$$

We see that P_0, P_1, and P_2 generate translations in the t, x, and y directions, respectively; D corresponds to dilations, B_1 and B_2 to Galilei boosts in the x and y directions, respectively. Finally R_0 corresponds to a rotation in the (u,v) plane, i.e., a constant change of phase of $\Psi(x,y,t)$ and R_1 to a change of phase of Ψ, linear in t, accompanied by constant shift in w (see below).

B. Loop structure of the DS symmetry algebra

Similarly as the algebra of the Kadomtsev–Petviashvili equation,[7] the DS symmetry algebra for $\delta_1=-\epsilon_1$ (and only in this case) can be embedded into a Kac–Moody type loop algebra.[13] To see this, let us restrict f, g, h, and l to be Laurent polynomials in t. A basis for this algebra is provided by the operators

$$\begin{aligned} X(t^n) &= t^n\,\partial_t+nt^{n-1}\Delta/2-n(n-1)t^{n-2}A_1/4 \\ &\qquad -n(n-1)(n-2)t^{n-3}W_1/4\,, \\ Y(t^n) &= t^nX-nt^{n-1}A_2/2-n(n-1)t^{n-2}W_2/2\,, \\ Z(t^n) &= t^nY-\epsilon_1nt^{n-1}A_3/2-\epsilon_1n(n-1)t^{n-2}W_3/2\,, \\ W(t^n) &= t^nA_4+nt^{n-1}W_4\,, \end{aligned} \tag{2.10}$$

where we have introduced the notation

$$\begin{aligned} &\Delta=x\partial_x+y\partial_y-u\partial_u-v\partial_v-2w\partial_w\,, \\ &X=\partial_x,\quad Y=\partial_y \\ &A_1=\tfrac{1}{2}(x^2+\epsilon_1y^2)(v\,\partial_u-u\,\partial_v),\quad A_2=x(v\,\partial_u-u\,\partial_v), \\ &A_3=y(v\,\partial_u-u\,\partial_v),\quad A_4=v\,\partial_u-u\,\partial_v, \\ &W_1=\tfrac{1}{2}(x^2+\epsilon_1y^2)\partial_w,\quad W_2=x\,\partial_w, \\ &W_3=y\,\partial_w,\quad W_4=\partial_w. \end{aligned} \tag{2.11}$$

The operators (2.11) form the basis of an 11-dimensional solvable Lie algebra. It has a ten-dimensional nilpotent ideal, the nilradical $\mathrm{NR}(L)=\{X,Y,A_1,A_2,A_3,A_4,W_1,W_2,W_3,W_4\}$. In turn the algebra $\mathrm{NR}(L)$ has an eight-dimensional uniquely defined maximal Abelian ideal $\{A_i\,W_i,\ i=1,\ldots,4\}$. The algebra (2.11) can be embedded into the simple Lie algebra $\mathrm{sl}(7,\mathbb{C})$. Indeed, consider the $\mathrm{sl}(7,\mathbb{C})$ matrix

$$\left\{\begin{matrix} \delta & 0 & x+\sqrt{-\epsilon_1}y & 0 & w_2+\epsilon_1(-\epsilon_1)^{1/2}w_3 & 0 & -2w_4 \\ 0 & -\delta & 0 & x+\sqrt{-\epsilon_1}y & a_2+\epsilon_1(-\epsilon_1)^{1/2}a_3 & 0 & -2a_4 \\ 0 & 0 & 2\delta & 0 & cw_1 & 0 & -w_2+\epsilon_1(-\epsilon_1)^{1/2}w_3 \\ 0 & 0 & 0 & 0 & ca_1 & 0 & a_2+\epsilon_1(-\epsilon_1)^{1/2}a_3 \\ 0 & 0 & 0 & 0 & -2\delta & 0 & x-(-\epsilon_1)^{1/2}y \\ 0 & 0 & 0 & 0 & 0 & \delta & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -\delta \end{matrix}\right\}. \tag{2.12}$$

Setting all entries but one equal to 0 and the remaining one equal to 1, we obtain 11 matrices having the same commutation relations as the vector fields (2.11) [$\delta = 1$ corresponds to the operator Δ, $x = 1$ or $y = 1$ to X or Y, respectively, $a_i = 1$ or $w_i = 1$ to A_i or W_i, respectively ($i = 1,\ldots,4$)]. This embedding provides us with an identification of the algebra generated by $X(t^n)$, $Y(t^n)$, $Z(t^n)$, and $W(t^n)$ of (2.10). We have obtained an infinite-dimensional subalgebra of the affine loop algebra,

$$\widehat{\mathrm{sl}}(7,\mathbb{C}) = \left\{[\mathbb{R}(t,t^{-1})\otimes \mathrm{sl}(7,\mathbb{C})]\oplus R(t,t^{-1})\frac{d}{dt}\right\}. \tag{2.13}$$

The vector fields $X(t^n)$ form a simple subalgebra isomorphic to the Z-graded algebra $\{R(t,t^{-1})d/dt\}$, which is in turn isomorphic to the Virasoro algebra (without a central extension).[13] Notice also that each element of (2.10) has a well-defined degree in a natural grading obtained by attributing the degree n to the monomial t^n and the degree μ ($0 \leqslant \mu \leqslant 6$) to each element of (2.11), where μ is the distance from the diagonal in the matrix (2.12) to the corresponding element ($\mu = 0$ for Δ, $\mu = 1$ for A_1, $\mu = 2$ for X, Y, and W_1, etc.). The degrees of $X(t^n)$, $Y(t^n)$, $Z(t^n)$, and $W(t^n)$ are thus $n-1$, $n+2$, $n+2$, and $n+5$, respectively.

C. The group transformations

The elements of the connected part of the symmetry group of the DSE's are obtained by integrating the general element of the DS Lie algebra (2.4). We consider separately the cases $f(t) = 0$ and $f(t) \neq 0$. In each case we write the vector field V in the form (2.2), where η_i and ϕ_i must be specified, and integrate the equations

$$\frac{d\tilde{x}}{d\lambda} = \eta_1, \quad \frac{d\tilde{y}}{d\lambda} = \eta_2, \quad \frac{d\tilde{t}}{d\lambda} = \eta_3,$$
$$\frac{d\tilde{u}}{d\lambda} = \phi_1, \quad \frac{d\tilde{v}}{d\lambda} = \phi_2, \quad \frac{d\tilde{w}}{d\lambda} = \phi_3, \tag{2.14}$$

where

$$\eta_i = \eta_i(\tilde{x},\tilde{y},\tilde{t},\tilde{u},\tilde{v},\tilde{w}), \quad \phi_i = \phi_i(\tilde{x},\tilde{y},\tilde{t},\tilde{u},\tilde{v},\tilde{w}).$$

The boundary conditions are

$$\tilde{x}|_{\lambda=0} = x, \quad \tilde{y}|_{\lambda=0} = y, \quad \tilde{t}\,|_{\lambda=0} = t,$$
$$\tilde{u}|_{\lambda=0} = u, \quad \tilde{v}|_{\lambda=0} = v, \quad \tilde{w}|_{\lambda=0} = w. \tag{2.15}$$

The results of this integration are presented below. For each of the two cases mentioned above, we give the transformed variables and the expression for the new solution in terms of the original one.

(i) Case $f(t) = 0$,

$$\tilde{x}(\lambda) = x + \lambda g(t), \quad \tilde{y}(\lambda) = y + \lambda h(t), \quad \tilde{t}(\lambda) = t,$$
$$\tilde{\Psi}(\tilde{x},\tilde{y},\tilde{t}) = \Psi(\tilde{x} - \lambda g(\tilde{t}),\tilde{y} - \lambda h(\tilde{t}),\tilde{t})\exp i\{(\lambda/2)g'(\tilde{t})[\tilde{x} - (\lambda/2)g(\tilde{t})]$$
$$+ (\epsilon_1/2)\lambda h'(\tilde{t})[\tilde{y} - (\lambda/2)h(\tilde{t})] - \lambda m(\tilde{t})\}, \tag{2.16}$$
$$\tilde{w}(\tilde{x},\tilde{y},\tilde{t}) = w[\tilde{x} - \lambda g(\tilde{t}),\tilde{y} - \lambda h(\tilde{t}),\tilde{t})] - (\lambda/2)g''(\tilde{t})[\tilde{x} - (\lambda/2)g(\tilde{t})] - (\epsilon_1/2)\lambda h''(\tilde{t})[\tilde{y} - (\lambda/2)h(\tilde{t})] + \lambda m'(\tilde{t}).$$

Setting $g(t) = h(t) = 0$ in (2.16), we see that the presence of $W(m)$ in the symmetry algebra simply means that the DSE's are invariant under an arbitrary time dependent change in the phase of Ψ, compensated by an appropriate transformation of w.

(ii) Case $f(t) \neq 0$,

$$G(t,\tilde{t}) = f^{1/2}(t)\int_t^{\tilde{t}} g(s)f^{-3/2}(s)ds, \quad H(t,\tilde{t}) = f^{1/2}(t)\int_t^{\tilde{t}} h(s)f^{-3/2}(s)ds, \quad \phi'(t) = \frac{1}{f(t)} \tag{2.17}$$

[$\phi(t)$ can be any antiderivative of $1/f(t)$], we have

$$\tilde{x}(\lambda) = [x + G(t,\tilde{t}(\lambda))][f(\tilde{t}(\lambda))/f(t)]^{1/2}, \quad \tilde{y}(\lambda) = [y + H(t,\tilde{t}(\lambda))][f(\tilde{t}(\lambda))/f(t)]^{1/2}, \quad \tilde{t}(\lambda) = \phi^{-1}(\phi(t) + \lambda),$$

$$\tilde{\Psi}(\tilde{x},\tilde{y},\tilde{t}) = \left[\frac{f(t)}{f(\tilde{t})}\right]^{1/2}\tilde{\Psi}(x,y,t)\exp i\left\{(\tilde{x}^2 + \epsilon_1\tilde{y}^2)\left\{\frac{[f'(\tilde{t}) - f'(t)]}{8f(\tilde{t})}\right\}\right.$$
$$+ \frac{1}{2}\tilde{x}[f(t)f(\tilde{t})]^{-1/2}\left[g(\tilde{t})\left[\frac{f(t)}{f(\tilde{t})}\right]^{1/2} - g(t) + \frac{1}{2}f'(t)G(t,\tilde{t})\right]$$
$$+ \frac{\epsilon_1}{2}\tilde{y}[f(t)f(\tilde{t})]^{-1/2}\left[h(\tilde{t})\left[\frac{f(t)}{f(\tilde{t})}\right]^{1/2} - h(t) + \frac{1}{2}f'(t)H(t,\tilde{t})\right]$$

$$
\begin{aligned}
&+\frac{1}{2f(t)}[g(t)G(t,\tilde{t})+\epsilon_1 h(t)H(t,\tilde{t})]-\frac{f'(t)}{8f(t)}[G^2(t,\tilde{t})+\epsilon_1 H^2(t,\tilde{t})]\\
&-\int_t^{\tilde{t}}\frac{g^2(s)+\epsilon_1 h^2(s)+2m(s)f(s)}{2f^2(s)}\,ds\Big\}\,,\\
\tilde{w}(\tilde{x},\tilde{y},\tilde{t}) = {}&[f(t)/f(\tilde{t})]w(x,y,t)\\
&-\frac{1}{8}(\tilde{x}^2+\epsilon_1\tilde{y}^2)f^{-2}(\tilde{t})\Big[f''(\tilde{t})f(\tilde{t})-\frac{1}{2}[f'(\tilde{t})]^2-f''(t)f(t)+\frac{1}{2}[f'(t)]^2\Big]\\
&-\frac{1}{4}\tilde{x}f^{-1/2}(t)f^{-3/2}(\tilde{t})\Big\{[2g'(\tilde{t})f(\tilde{t})-g(\tilde{t})f'(\tilde{t})]\Big[\frac{f(t)}{f(\tilde{t})}\Big]^{1/2}\\
&-[2g'(t)f(t)-g(t)f'(t)]+\Big[f''(t)f(t)-\frac{1}{2}[f'(t)]^2\Big]G(t,\tilde{t})\Big\}\\
&-\frac{\epsilon_1}{4}\tilde{y}f^{-1/2}(t)f^{-3/2}(\tilde{t})\Big\{[2h'(\tilde{t})f(\tilde{t})-h(\tilde{t})f'(\tilde{t})]\Big[\frac{f(t)}{f(\tilde{t})}\Big]^{1/2}\\
&-[2h'(t)f(t)-h(t)f'(t)]+\Big[f''(t)f(t)-\frac{1}{2}[f'(t)]^2\Big]H(t,\tilde{t})\Big\}\\
&-\frac{1}{4f(t)f(\tilde{t})}\{[2g'(t)f(t)-g(t)f'(t)]G(t,\tilde{t})+\epsilon_1[2h'(t)f(t)-h(t)f'(t)]H(t,\tilde{t})\}\\
&+\frac{1}{8f(t)f(\tilde{t})}\Big[f''(t)f(t)-\frac{1}{2}[f'(t)]^2\Big][G^2(t,\tilde{t})+\epsilon_1 H^2(t,\tilde{t})]\\
&+\frac{g^2(\tilde{t})+\epsilon_1 h^2(\tilde{t})+4m(\tilde{t})f(\tilde{t})}{4f^2(\tilde{t})}-\frac{g^2(t)+\epsilon_1 h^2(t)+4m(t)f(t)}{4f(t)f(\tilde{t})}\,.
\end{aligned}
\tag{2.18}
$$

The variables x, y, and t appearing on the right-hand side of the expressions for Ψ and w in (2.18) are to be interpreted as functions of $\tilde{x}$, $\tilde{y}$, and $\tilde{t}$, i.e.,

$$
\begin{aligned}
&x=[\tilde{x}+G(\tilde{t},t)][f(t)/f(\tilde{t})]^{1/2},\\
&y=[\tilde{y}+H(\tilde{t},t)][f(t)/f(\tilde{t})]^{1/2},\quad t=\phi^{-1}(\phi(\tilde{t})-\lambda)\,.
\end{aligned}
\tag{2.19}
$$

Note that by construction, a one-dimensional Lie subgroup of transformations is generated if one fixes the functions g, h, l in (2.16), or f, g, h, l in (2.18), and then allows the parameter λ to take on arbitrary real values.

The expressions for $\tilde{\Psi}$ and $\tilde{w}$ in (2.16) and (2.18) can be used to generate new solutions of the DSE's from known ones. More precisely, if (Ψ,w) is a local solution of the DSE's in the neighborhood of (x,y,t) then $(\tilde{\Psi},\tilde{w})$ given by (2.16) or (2.18) will be a local solution of the DSE's in the neighborhood of $(\tilde{x}(\lambda),\tilde{y}(\lambda),\tilde{t}(\lambda))$, In particular, the application of the transformations (2.16) and (2.18) to the "trivial" constant solution

$$
\Psi(x,y,t)=\Psi_0,\quad w(x,y,t)=-\epsilon_1|\Psi_0|^2\,,
\tag{2.20}
$$

provides us with a family of solutions depending on three arbitrary functions of time in the case of (2.16) and four arbitrary functions of time in the case of (2.18).

By introducing new functions of time, it is possible to obtain a much more simple expression than (2.18) for the elements of the symmetry group of the DSE's when $f(t)\neq 0$. Let $a(t)$, $b(t)$, $c(t)$, and $e(t)$ be arbitrary real-valued functions of class C^∞ with the restriction that $a(t)\neq 0$. Then, for a fixed value of the parameter λ_0, take $f(t)$, $g(t)$, $h(t)$, and $l(t)$ in (2.17) and (2.18) to be the solutions of the following system of functional equations:

$$
\begin{aligned}
&a(\tilde{t})=f(\tilde{t})/f(t),\\
&\frac{b(\tilde{t})}{a(\tilde{t})}=[f(t)f(\tilde{t})]^{-1/2}\Big\{g(\tilde{t})\Big[\frac{f(t)}{f(\tilde{t})}\Big]^{1/2}\\
&\qquad-g(t)+\frac{1}{2}f'(t)G(t,\tilde{t})\Big\},\\
&\frac{c(\tilde{t})}{a(\tilde{t})}=[f(t)f(\tilde{t})]^{-1/2}\Big\{h(\tilde{t})\Big[\frac{f(t)}{f(\tilde{t})}\Big]^{1/2}\\
&\qquad-h(t)+\frac{1}{2}f'(t)H(t,\tilde{t})\Big\},\\
&\int_t^{\tilde{t}}\frac{b^2(s)+\epsilon_1c^2(s)+2e(s)a(s)}{a^2(s)}\,ds\\
&\quad=\int_t^{\tilde{t}}\frac{g^2(s)+\epsilon_1h^2(s)+2m(s)f(s)}{f^2(s)}\,ds\\
&\qquad-\frac{1}{f(t)}[g(t)G(t,\tilde{t})+\epsilon_1h(t)H(t,\tilde{t})]\\
&\qquad+\frac{f'(\tilde{t})}{4f(t)}[G^2(t,\tilde{t})+\epsilon_1H^2(t,\tilde{t})]\,,\\
&\qquad\tilde{t}=\phi^{-1}(\phi(t)+\lambda_0),\quad \phi'(t)=1/f(t)\,.
\end{aligned}
\tag{2.21}
$$

Under these conditions, it can be verified that for appropriate constants of integration c_1, c_2, c_3, and c_4, the transformations (2.18) reduce to the following transformations when $\lambda=\lambda_0$ (and only for this value of λ):

$$
\begin{aligned}
&\tilde{x}=\Big[x+\int b(\tilde{t})a^{-3/2}(\tilde{t})d\tilde{t}+c_1\Big]a^{1/2}(\tilde{t}),\\
&\tilde{y}=\Big[y+\int c(\tilde{t})a^{-3/2}(\tilde{t})d\tilde{t}+c_2\Big]a^{1/2}(\tilde{t}),
\end{aligned}
$$

$$\tilde{t}=\zeta^{-1}(t),\quad \zeta(t)=\int\frac{dt}{a(t)}+c_3,$$

$$\tilde{\Psi}(\tilde{x},\tilde{y},\tilde{t})=a^{-1/2}\Psi(x,y,t)\exp i\left\{\frac{a'}{8a}(\tilde{x}^2+\epsilon_1\tilde{y}^2)+\frac{b}{2a}\tilde{x}\right.$$
$$\left.+\frac{\epsilon_1 c}{2a}\tilde{y}-\frac{1}{2}\int\frac{b^2+\epsilon_1c^2+2ea}{a^2}\,d\tilde{t}-c_4\right\},\tag{2.22}$$
$$\tilde{w}(\tilde{x},\tilde{y},\tilde{t})=a^{-1}w(x,y,t)-\frac{1}{8}\left(\frac{aa''-\frac{1}{2}a'^2}{a^2}\right)(\tilde{x}^2+\epsilon_1\tilde{y}^2)$$
$$-\frac{1}{4}\left(\frac{2b'a-ba'}{a^2}\right)\tilde{x}-\frac{\epsilon_1}{4}\left(\frac{2c'a-ca'}{a^2}\right)\tilde{y}$$
$$+\frac{1}{4}\left(\frac{b^2+\epsilon_1c^2+4ea}{a^2}\right).$$

The functions a, b, c, e, and the derivatives of these functions appearing in the right-hand side of the expressions for $\tilde{\Psi}$ and $\tilde{w}$ in (2.22) are all evaluated at $\tilde{t}$. Moreover, the variables x, y, t appearing in the argument of Ψ and w in the same expressions are to be interpreted as functions of $\tilde{x}$, $\tilde{y}$, $\tilde{t}$, i.e.,

$$x=\tilde{x}a^{-1/2}(\tilde{t})-\int b(\tilde{t})a^{-3/2}(\tilde{t})d\tilde{t}-c_1,$$
$$y=\tilde{y}a^{-1/2}(\tilde{t})-\int c(\tilde{t})a^{-3/2}(\tilde{t})d\tilde{t}-c_2,\tag{2.23}$$
$$t=\zeta(\tilde{t}).$$

It should be pointed out that the constants c_1, c_2, c_3, and c_4 can be omitted when using (2.22) since they can always be removed by applying the transformation $\exp\{-X(c_3)-Y(c_1)-Z(c_2)-W(c_4)\}$ prior to the application of (2.22).

Finally, let us mention that the DSE's (1.1) are also invariant under a group of discrete transformations, generated by the transformations

$$\begin{array}{lllll}
x\to -x, & y\to y, & t\to t, & \Psi\to\Psi, & w\to w,\\
x\to x, & y\to -y, & t\to t, & \Psi\to\Psi, & w\to w,\\
x\to x, & y\to y, & t\to t, & \Psi\to -\Psi, & w\to w,\\
x\to x, & y\to y, & t\to -t, & \Psi\to\Psi^*, & w\to w.
\end{array}\tag{2.24}$$

III. ONE- AND TWO-DIMENSIONAL SUBALGEBRAS OF THE DAVEY–STEWARTSON SYMMETRY ALGEBRA

In order to perform symmetry reduction for the DSE's in a systematic manner, we need to know all subgroups of the symmetry group having generic orbits of codimension 1 and 2 in the $\{x,y,t\}$ space. This is equivalent to performing a classification of all one- and two-dimensional subalgebras of the DS algebra into conjugacy classes under the adjoint action of the DS group, i.e., the group leaving the equations invariant.

The method is exactly the same as the one employed recently for the Kadomtsev–Petviashvili equation,[7] and is an adaptation of methods developed earlier for classifying subalgebras of finite-dimensional Lie algebras.[18,19]

The first step is to classify subalgebras of the factor algebra $S=\{X(F)\}$ in the Levi decomposition (2.8) For this we can use results obtained earlier[7] for an isomorphic algebra. Thus every nontrivial one-dimensional subalgebra of S is conjugate to $\{X(1)\}$ and every two-dimensional subalgebra to $\mathrm{aff}(1,\mathbf{R})=\{X(1),X(t)\}$.

One-dimensional subalgebras of the entire DS algebra will thus have the form $\{X(1)+Y(g)+Z(h)+W(m)\}$, or $\{Y(g)+Z(h)+W(m)\}$. Using the transformations (2.16)–(2.19) we can show that every subalgebra of the first type is conjugate to $X(1)$.

The subalgebras of the second type split into several classes depending on which of the functions $g(t)$, $h(t)$, and $m(t)$ are nonzero (in the considered t interval). We drop all details and present representatives of each conjugacy class of one-dimensional subalgebras of the DS algebra in Table I. The classification is under the entire DS group including the discrete transformations (2.24).

In column 1 we introduce a name for each class of subalgebras. In column 2 we give the basis element for each representative subalgebra. In column 3 we present the normalizer of each subalgebra in the DS algebra, i.e., the maximal subalgebra $L_0\subset L$ satisfying

$$[X,X_0]=\lambda X_0,\quad X\in L_0,\tag{3.1}$$

where $\lambda\in\mathbf{R}$ is a constant and X_0 is the corresponding basis element in column 2. In column 4 we give the conditions

TABLE I. One-dimensional subalgebras of the Davey–Stewartson algebra ($a>0$ and $\lambda\in\mathbf{R}$ are constants, h, F, H, G, and L are functions of t).

No.	Basis element	Normalizer	Characterization of conjugacy class
$L_{1,1}$	$X(1)$	$X(t),\ X(1),\ Y(1),\ Z(1),\ W(1)$	$f\neq 0$
$L^a_{1,2}$	$Y(1)+aZ(1)$	$X(t),\ X(1),\ Y(-\epsilon_1 aH)+Z(H)$ $Y(1),\ Z(1),\ W(L)$	$f=0,\ h=\pm ag\neq 0$
$L_{1,3}(h)$ $h'\neq 0$	$Y(1)+Z(h)$	$-\epsilon_1 Y[\int_0^t(hH'-h'H)ds]+Z(H),$ $Y(1),\ Z(h),\ W(L)$	$f=0,\ g\neq 0$ $h\neq\lambda g$
$L_{1,4}$	$Z(1)$	$X(t),\ X(1),\ Y(G),\ Z(1),\ W(L)$	$f=g=0,\ h\neq 0$
$L_{1,5}$	$W(t)$	$X(t),\ Y(G),\ Z(H),\ W(L)$	$f=g=h=0,$ $m\neq 0$
$L_{1,6}$	$W(1)$	$X(F),\ Y(G),\ Z(H),\ W(L)$	$f=g=h=0$ $m=\lambda\neq 0$

under which a general element of the form (2.4) can be transformed into a constant multiple of the element in column 2.

Two isomorphy classes of two-dimensional Lie algebras $\{X_1,X_2\}$ exist, Abelian ($2A_1$) and non-Abelian (A_2), with commutation relation

$$[X_1,X_2]=0 \quad \text{or} \quad [X_1,X_2]=X_1\,, \tag{3.2}$$

respectively. To obtain all such algebras we let X_1 run through all the standard forms of Table I. The other element X_2 must then lie in the normalizer nor$\{X_1\}$ and can be further simplified using the Lie group Nor$\{X_1\}$ corresponding to the algebra nor$\{X_1\}$.

The results are summarized in Table II. Certain redundancies have been left in Tables I and II. Thus two one-dimensional subalgebras $L_{1,3}(h_1)$ and $L_{1,3}(h_2)$ are conjugate to each other if there exist two constants λ and μ, such that

$$h_2(t)=h_1(\lambda t+\mu)\,. \tag{3.3}$$

Similar redundancies exist in Table II and can be removed, e.g., by fixing the values of the function $h(t)$ and its derivative at some point $t=t_0$. Since this has no consequences for symmetry reduction, we shall not dwell on it here.

IV. SYMMETRY REDUCTION FOR THE DAVEY-STEWARTSON EQUATIONS

We shall now use the results of the previous sections to reduce the DSE's to a system of equations involving two independent variables only. To do this we make use of the one-dimensional subalgebras of the DS algebra, listed in Table I. The method is standard and quite simple. We consider an auxiliary function $F(x,y,t,u,v,w)$ and request that it be annihilated by the elements of the one-dimensional subalgebra $\{X\}$:

$$XF=0. \tag{4.1}$$

TABLE II. Two-dimensional subalgebras of the DS algebra ($a\geqslant 0$, $b\in\mathbf{R}$, $k=0$, are constants, h, H, and m are functions of t).

No.	Type	Basis element
$L_{2,1}^{a,k}$	$2A_1$	$X(1),\ Y(1)+aZ(1)+kW(1)$
$L_{2,2}^{k}$	$2A_1$	$X(1),\ Z(1)+kW(1)$
$L_{2,3}$	$2A_1$	$X(1),\ W(1)$
$L_{2,4}^{a,b,h}$	$2A_1$	$Y(1)+aZ(1),\ Y(-\epsilon_1 ah)+Z(h)+bZ(1)$
$L_{2,5}^{a,h,m}$	$2A_1$	$Y(1)+aZ(1),\ Y(-\epsilon_1 ah)+Z(h)+W(m)$ $[m=0 \text{ if } (a^2,\epsilon_1)\neq(1,-1)]$
$L_{2,6}^{h,H}$, $h'\neq 0$	$2A_1$	$Y(1)+Z(h),\ -\epsilon_1 Y[\int_0^t(hH'-h'H)ds]+Z(H)$
$L_{2,7}^{h,m}$, $h'\neq 0$	$2A_1$	$Y(1)+Z(h),\ W(m)$
$L_{2,8}^{m}$	$2A_1$	$Z(1),\ W(m)$
$L_{2,9}^{m,}$	$2A_1$	$W(t),\ W(m)$
$L_{2,10}^{k}$	A_2	$X(1),\ X(t)+kW(1)$
$L_{2,11}^{a,}$	A_2	$Y(1)+aZ(1),\ 2X(t)$
$L_{2,12}$	A_2	$Z(1),\ 2X(t)$
$L_{2,13}$	A_2	$W(t),\ -X(t)$

Equation (4.1) implies that F is a function of five variables only, namely the invariants of the Lie group generated by X. Two invariants ξ and η can be chosen to depend on x, y, and t only, these are the new symmetry variables. The remaining invariants yield the dependence of u, v, and w (i.e., ψ and w) on the symmetry variables.

Only vector fields involving derivatives with respect to the independent variables yield reductions. Hence we shall only use the subalgebras $L_{1,1},\ldots,L_{1,4}$ of Table I. We shall perform the reduction using the "standard" basis elements of Table I. The result for a general vector field (2.4) is obtained from the results for a simplified one by applying a general group transformation (2.16)–(2.24).

A. The algebra $L_{1,1}$

The equation $X(1)F(x,y,t,u,v,w)=0$ tells us that the invariants of $\exp X(1)$ are x,y,u,v, and w. The reduction is hence obtained by setting

$$\begin{aligned}&\Psi(x,y,t)=\phi(\xi,\eta),\quad \xi=x,\quad \eta=y,\\ &w(x,y,t)=Q(\xi,\eta)\,.\end{aligned} \tag{4.2}$$

Substituting into the DSE's (1.1) we obtain the reduced system

$$\phi_{\xi\xi}+\epsilon_1\phi_{\eta\eta}=\epsilon_2|\phi|^2\phi+\phi Q\,, \tag{4.3a}$$

$$Q_{\xi\xi}+\delta_1 Q_{\eta\eta}=\delta_2(|\phi|^2)_{\eta\eta}\,. \tag{4.3b}$$

Applying a general DS group transformation to a solution of (4.3) we obtain a class of solutions of the DSE's, depending on four arbitrary functions $f(t)$, $g(t)$, $h(t)$, and $l(t)$. Thus assuming $f(t)\neq 0$, we obtain

$$\begin{aligned}\Psi&=\phi(\xi,\eta)f^{-1/2}\exp i\Big[\frac{1}{8}(x^2+\epsilon_1 y^2)\frac{f'}{f}\\ &\quad+\frac{1}{2f}(xg+\epsilon_1 yh)-\frac{1}{2}\int\frac{\epsilon_1 h^2+g^2+2mf}{f^2}ds\Big],\\ W&=Q(\xi,\eta)\frac{1}{f}-\frac{1}{8f^2}\Big(ff''-\frac{1}{2}f'^2\Big)(x^2+\epsilon_1 y^2)\\ &\quad-\frac{x}{4f^2}(2g'f-gf')-\frac{\epsilon_1 y}{4f^2}(2h'f-hf')\\ &\quad+\frac{1}{4}\frac{g^2+\epsilon_1 h^2+4mf}{f^2},\\ \xi&=xf^{-1/2}-\int_0^t g(s)[f(s)]^{-3/2}ds\,,\\ \eta&=yf^{-1/2}-\int_0^t h(s)[f(s)]^{-3/2}ds\,.\end{aligned} \tag{4.4}$$

Substituting (4.4) into the DSE's (1.1) we find that $\phi(\xi,\eta)$ and $Q(\xi,\eta)$ must satisfy Eqs. (4.3a) and

$$\begin{aligned}&8[Q_{\xi\xi}+\delta_1 Q_{\eta\eta}-\delta_2(|\phi|^2)_{\eta\eta}]\\ &\quad=(\delta_1\epsilon_1+1)[2ff_{tt}-(f_t)^2],\end{aligned} \tag{4.5}$$

which reduces to (4.3b) if $\delta_1=-\epsilon_1$ or if $f(t)=(a+bt)^2$ [see (2.6)].

B. The algebra $L^a_{1,2}$

The equation $[Y(1)+aZ(1)]F=0$ implies a reduction obtained by setting

$$\psi(x,y,t)=\Omega(\xi,\zeta),\quad w(x,y,t)=Q(\xi,\zeta),\qquad \xi=t,\quad \zeta=y-ax\,. \tag{4.6}$$

By substituting into the DS equations we obtain the reduced system

$$i\Omega_\xi+(a^2+\epsilon_1)\Omega_{\zeta\zeta}=\epsilon_2|\Omega|^2\Omega+Q\Omega\,, \tag{4.7a}$$

$$(a^2+\delta_1)Q_{\zeta\zeta}=\delta_2(|\Omega|^2)_{\zeta\zeta}\,. \tag{4.7b}$$

This system can be further simplified. We solve the second equation (choosing $a^2\neq-\delta_1$):

$$Q(\xi,\zeta)=[\delta_2/(a^2+\delta_1)]|\Omega|^2+\alpha(\xi)\zeta+\beta(\xi)\,, \tag{4.8}$$

where $\alpha(\xi)$ and $\beta(\xi)$ are arbitrary functions. Expression (4.8) can be substituted back into Eq. (4.7a) and we obtain an equation for $\Omega(\xi,\zeta)$ alone. A transformation of the dependent and independent variables can be found that takes (4.7a) into the nonlinear Schrödinger equation. The final result is

$$\begin{aligned}\Psi(x,y,t)&=\{\epsilon_4(a^2+\delta_1)/[\epsilon_2(a^2+\delta_1)+\delta_2]\}^{1/2}\phi(\xi,\eta)\\&\qquad\times\exp i[(y-ax)F(t)+G(t)]\,,\\w(x,y,t)&=\{\delta_2\epsilon_4/[\epsilon_2(a^2+\delta_1)+\delta_2]\}|\phi(\xi,\eta)|^2\\&\qquad+\alpha(t)(y-ax)+\beta(t)\,,\\F(t)&=-\int\alpha(t)dt,\\G(t)&=-\int[(a^2+\epsilon_1)F^2(t)+\beta(t)]dt\,,\\H(t)&=-2[\epsilon_3(a^2+\epsilon_1)]^{1/2}\int F(t)dt,\\\epsilon_4&=\operatorname{sgn}\frac{a^2+\delta_1}{\epsilon_2(a^2+\delta_1)+\delta_2}\,,\\\xi&=\epsilon_3t,\quad\eta=[\epsilon_3(a^2+\epsilon_1)]^{-1/2}(y-ax)+H(t),\\\epsilon_3&=\operatorname{sgn}(a^2+\epsilon_1)\,.\end{aligned} \tag{4.9}$$

Here $\alpha(t)$ and $\beta(t)$ are arbitrary functions of time, a is a constant, and $\phi(\xi,\eta)$ satisfies the nonlinear Schrödinger equation

$$i\phi_\xi+\phi_{\eta\eta}=\epsilon_3\epsilon_4\phi|\phi|^2\,. \tag{4.10}$$

We shall not present the more general solution, obtained by applying a general DS group element to the solution (4.9).

C. The algebra $L_{1,3}(h)$

We have

$$[Y(1)+Z(h)]F=\{\partial_x+h\,\partial_y-(\epsilon_1/2)y[h'(v\,\partial_u-u\,\partial_v)+h''\,\partial_w]\}F=0\,. \tag{4.11}$$

The characteristic system for (4.11) is

$$\frac{dx}{1}=\frac{dy}{h}=-\frac{2\,du}{\epsilon_1yh'v}=\frac{2\,dv}{\epsilon_1yh'u}=\frac{2\,dw}{\epsilon_1yh''}\,. \tag{4.12}$$

By solving (4.12) we obtain

$$\begin{aligned}\Psi&=\phi(\xi,\eta)\exp(i(\epsilon_1/4)(h'/h)y^2),\quad\xi=t\,,\\W&=Q(\xi,\eta)-(\epsilon_1/4)(h''/h)y^2,\quad\eta=y-h(t)x\,,\end{aligned} \tag{4.13}$$

where the DSE's imply

$$i\phi_\xi+(\epsilon_1+h^2)\phi_{\eta\eta}+\frac{ih'}{h}\eta\phi_\eta+\frac{i}{2}\frac{h'}{h}\phi=\epsilon_2|\phi|^2\phi+\phi Q\,, \tag{4.14a}$$

$$(h^2+\delta_1)Q_{\eta\eta}-\frac{\epsilon_1\delta_1}{2}\frac{h''}{h}=\delta_2(|\phi|^2)_{\eta\eta}\,. \tag{4.14b}$$

The system (4.14) can be further simplified. Solving (4.14b) and substituting into (4.14a), we find

$$Q=\frac{\delta_2}{h^2+\delta_1}|\phi|^2+\frac{\epsilon_1\delta_1}{4(h^2+\delta_1)}\frac{h''}{h}\eta^2+\alpha(t)\eta+\beta(t)\,, \tag{4.15}$$

$$\begin{aligned}&i\phi_\xi+(\epsilon_1+h^2)\phi_{\eta\eta}+\frac{ih'}{h}\eta\phi_\eta\\&\quad+\left(\frac{i}{2}\frac{h'}{h}-\frac{\epsilon_1\delta_1}{4(h^2+\delta_1)}\frac{h''}{h}\eta^2-\alpha\eta-\beta\right)\phi\\&\quad=\left(\epsilon_2+\frac{\delta_2}{h^2+\delta_1}\right)|\phi|^2\phi\,.\end{aligned} \tag{4.16}$$

Equation (4.16) can be reduced to a nonlinear Schrödinger equation with variable coefficients. To see this, set

$$Q=A(t)\Omega(\xi,\zeta)\exp[i(\eta^2H+\eta F+G)]\,,\qquad\zeta=\gamma(t)\eta+K(t)\,. \tag{4.17}$$

We choose $H(t)$ to satisfy a Riccati equation

$$H'+4(\epsilon_1+h^2)H^2+\frac{\epsilon_1\delta_1}{4(h^2+\delta_1)}\frac{h''}{h}+2\frac{h'}{h}H=0 \tag{4.18}$$

and the other functions in (4.17) to satisfy

$$\begin{aligned}&F'+4(\epsilon_1+h^2)HF+(h'/h)F+\alpha=0\,,\\&G'+(\epsilon_1+h^2)F^2+\beta=0\,,\\&A=h^{-1/2}\exp\left[-2\int(\epsilon_1+h^2)H\,dt\right],\\&K=-2\int(\epsilon_1+h^2)\gamma F\,dt\,,\quad\gamma=A^2.\end{aligned} \tag{4.19}$$

The function Ω in (4.17) then satisfies the equation

$$i\Omega_\xi+(\epsilon_1+h^2)A^4\Omega_{\xi\xi}=(\epsilon_2+\delta_2/(h^2+\delta_1))A^2|\Omega|^2\Omega\,. \tag{4.20}$$

For $\delta_1=-\epsilon_1$ a particular solution of the Riccati equation (4.18) is

$$H=h'/4(h^2-\epsilon_1)h\,. \tag{4.21}$$

D. The algebra $L_{1,4}$

The algebra generated by $Z(1)=\partial_y$ leads in a simple manner to the nonlinear Schrödinger equation. Indeed a straightforward reduction with $\psi=\Omega(x,t)$, $w=Q(x,t)$ yields

$$i\Omega_t + \Omega_{xx} = \epsilon_2 |\Omega|^2 \Omega + \Omega Q, \quad Q_{xx} = 0 . \tag{4.22}$$

Putting

$$\begin{aligned} &\Psi = \phi(t, \xi) e^{i[Fx + G]}, \\ &w = \alpha(t) x + \beta(t), \quad \xi = x + H(t) , \end{aligned} \tag{4.23}$$

with

$$F(t) = -\int \alpha(t) dt, \quad G(t) = -\int (F^2 + \beta) dt,$$

$$H(t) = -2\int F\,dt,$$

we find that $\phi(t,\xi)$ satisfies the nonlinear Schrödinger equation (1.2).

The algebras of Table II could be used to reduce the DSE's to various systems of nonlinear ordinary differential equations. These are easy to obtain and we shall not go into them here.

V. CONCLUSIONS

We have shown that the Davey–Stewartson equations (1.1) have an infinite-dimensional symmetry group. Moreover, for the integrable case when $\delta_1 = -\epsilon_1$ in (1.1), the symmetry Lie algebra has a loop algebra structure, similar to that of all other known integrable nonlinear differential equations in 2 + 1 dimensions.[7–12]

One-dimensional subalgebras of the symmetry algebra have been used in Sec. IV to reduce the DSE's to one of three two-dimensional systems. These are the system (4.3), the nonlinear Schrödinger equation (4.10) and Eq. (4.16). Large classes of solutions of the nonlinear Schrödinger equation are known (solitons, multisolitons, background radiation, quasiperiodic solutions).[4,5] The system (4.3) and Eq. (4.16) have, to our knowledge, not been studied in the literature. They merit a separate investigation and we plan to return to them in the future.

ACKNOWLEDGMENTS

The final version of this article was written while one of the authors (P.W.) was visiting the Institut de Physique Nucléaire, Université Claude Bernard, Lyon 1. He thanks Professor M. Kibler and Professor M. E. Elbaz for their hospitality.

This work was partially supported by research grants from the Natural Sciences and Engineering Research Council of Canada and the Fonds FCAR du Gouvernement du Québec.

[1] A. Davey and K. Stewartson, Proc. R. Soc. London Ser. A **338**, 101 (1974).

[2] V. V. Kadomtsev and V. I. Petviashvili, Sov. Phys. Dokl. **15**, 539 (1970).

[3] D. Anker and N. C. Freeman, Proc. R. Soc. London Ser. A **360**, 529 (1978).

[4] M. J. Ablowitz and H. Segur, *Solitons and the Inverse Scattering Transform* (SIAM, Philadelphia, 1981).

[5] S. P. Novikov, S. V. Manakov, L. P. Pitaevskii, and V. E. Zakharov, *Theory of Solitons. The Inverse Method* (Plenum, New York 1984).

[6] M. J. Ablowitz and A. S. Fokas, in *Nonlinear Phenomena, Lecture Notes in Physics*, Vol. 189 (Springer, New York, 1983).

[7] D. David, N. Kamran, D. Levi, and P. Winternitz, Phys. Rev. Lett. **55**, 2111 (1985); J. Math. Phys. **27**, 1225 (1986).

[8] B. Champagne and P. Winternitz, Montréal, CRM 1278 preprint, 1985.

[9] B. Dorizzi, B. Grammaticos, A. Ramani, and P. Winternitz, J. Math. Phys. **27**, 2848 (1986).

[10] R.A. Leo, L. Martina, and G. Soliani, J. Math. Phys. **27**, 2623 (1986).

[11] L. Martina and P. Winternitz, Montreal, CRM 1987, to be published.

[12] D. David, D. Levi, and P. Winternitz, Phys. Lett. A **118**, 390 (1986).

[13] V. Kac, *Infinite Dimensional Lie Algebras* (Birkhäuser, Boston, 1984).

[14] M. Jimbo and T. Miwa, Publ. Res. Inst. Math. Sci. Kyoto Univ. **19**, 943 (1983).

[15] P. J. Olver, *Applications of Lie Groups to Differential Equations* (Springer, New York, 1986).

[16] F. Schwarz, Comp. Phys. Commun. **27**, 179 (1982); Computing **34**, 91 (1985).

[17] N. Jacobson, *Lie Algebras* (Wiley–Interscience, New York, 1962).

[18] J. Patera, P. Winternitz, and H. Zassenhaus, J. Math, Phys. **16**, 1597, 1615 (1975).

[19] J. Patera, R. T. Sharp, P. Winternitz, and H. Zassenhaus, J. Math. Phys. **18**, 2259 (1977).

SL(3,R) as the group of symmetry transformations for all one-dimensional linear systems

M. Aguirre
Instituto de Física, Universidad Católica de Valparaíso, Casilla 4059, Valparaíso, Chile

J. Krause
Facultad de Física, Pontificia Universidad Católica de Chile, Casilla 6177, Santiago 22, Chile

(Received 30 January 1987; accepted for publication 26 August 1987)

The converse problem of similarity analysis is solved in general for the finite symmetry transformations of any inhomogeneous ordinary linear differential equation of the second order $\ddot{x} + f_2(t)\dot{x} + f_1(t)x = f_0(t)$. The eight-parameter realizations of the symmetry group are obtained in the form $\mathscr{F}^{-1}\mathscr{P}_2\mathscr{F}$, where $\mathscr{F}$ stands for transformations of (t,x) that depend exclusively on the fundamental solutions of the equation, and where $\mathscr{P}_2$ is an arbitrary projective transformation in the plane. Thus it is shown that the full point symmetry group corresponds to SL(3,R) indeed, without recourse to the Lie algebra. Also, a technique is obtained for calculating the finite point symmetry realization of SL(3,R) for any given one-dimensional linear system. Some miscellaneous examples are given.

I. INTRODUCTION

In this paper we are interested in the point symmetry properties of one-dimensional linear systems in Newtonian mechanics. It is our aim to give a *unified* treatment of the similarity properties of such systems, in order to show that SL(3,R) is the maximal group of point symmetry transformations for *all* linear inhomogeneous ordinary differential equations of the second order, in one real dependent variable.[1] To this end, we shall use the group elements (instead of the Lie algebra generators[1]) to uncover the symmetry group. Furthermore, as a striking feature of this approach, one obtains a technique for calculating the *realization* of SL(3,R), as a group of point symmetry transformations, for any given second-order linear differential equation in one dimension.

In the last few years there has been considerable progress in the study of symmetries and invariants in classical mechanics.[2] Different approaches to this subject are found in contemporary literature, which start from different concepts of what is the basic dynamical formulation for studying the symmetries of mechanical systems in general.[3] Thus Noether and Lie symmetries have been distinguished,[3] and Noether and non-Noether constants of motion[4] have been discussed in the recent literature. As a matter of fact, the overriding lesson seems to be that all these approaches are equally fruitful for the theory of symmetry in dynamics.[3]

Interesting progress on this subject has been made in recent years from the standpoint of continuous groups of transformations of the equations of motion.[5] It has been found that the demand of invariance of equations of motion yields not only the conventional conservation laws, but also the "accidental symmetries" and the corresponding conservation laws.[6] In particular, let us recall that the maximal Lie group of point symmetry transformations for the simple harmonic oscillator was identified by Wulfman and Wybourne[7] as the group SL(3,R).[8] Wulfman and Wybourne, however, present the space-time realization of SL(3,R) (for the case $\ddot{x} + \omega^2 x = 0$) only through its *one-parameter subgroups*. Hence their realization of the elements of the group is not complete, because the *analytic continuation* of the one-parameter space-time transformations over the group manifold is still missing in that work. This analytic continuation being necessary, of course, in order to have a complete space-time realization of the generic element of SL(3,R) with the eight parameters included.

In a previous paper,[9] the problem set by the space-time realization of the general element of SL(3,R) was revisited for the case of the harmonic oscillator. In that paper, the general realization of the group was calculated with the eight parameters included; and the group was shown to be SL(3,R), without recourse to the Lie algebra. Therefore a *synthetic method* was adopted, considering SL(3,R) as a group of space-time automorphisms that interconverts one admissible world line of the oscillator into another. In this manner, it can be found that the point symmetry group of $\ddot{x} + \omega^2 x = 0$ becomes *faithfully* realized as the *projective group in the plane*. In fact, it is well known that SL(3,R) and the projective group in the plane are isomorphic (cf., also, *infra*). The important point to remark concerning the synthetic method lies in its *linear* character, which rests exclusively on the property that $\ddot{x} + \omega^2 x = 0$ is a *linear* differential equation of the *second order*. Indeed, in the present paper we shall take advantage of this fact, extending the synthetic method to study the point symmetries of *any* linear second-order differential equation. In this fashion, we shall obtain complete *generalization* and *unification* of a great amount of work, which has been previously performed on the symmetry properties of one-dimensional linear systems in classical mechanics.

The organization of this paper is as follows. We first briefly examine some features of the finite symmetry analysis of a one-dimensional linear system (Sec. II). Then we tackle the general converse similarity problem for one-dimensional linear systems by means of a new approach (Sec. III), which reveals the central role played by the projective group. In Sec. IV, we establish the relation with SL(3,R). Finally, Sec. V contains some examples of the formalism.

0022-2488/88/010009-07$02.50

II. SOME REMARKS ON THE FINITE POINT SYMMETRY TRANSFORMATIONS OF $\ddot{x}+f_2\dot{x}+f_1x=f_0$

The standard form of the linear differential equation of the second order will be taken to be

$$\ddot{x}+f_2(t)\dot{x}+f_1(t)x=f_0(t), \tag{2.1}$$

where all variables and functions are real. In the sequel it will be assumed that we are working within an interval of the independent variable, $t_1<t<t_2$, throughout which $f_1(t)$ and $f_2(t)$ [as well as $f_0(t)$] are continuous one-valued functions; i.e., there exists a unique continuous solution $x=w(t)$, within $t_1<t<t_2$.

For our purposes we may consider the whole space-time of the system as decomposed in continuous bands, say $\{t_\alpha<t<t_{\alpha+1}, -\infty<x<+\infty\}$, $\alpha=1,2,\ldots$, within each of which the problem set by Eq. (2.1) is well posed and the fundamental existence theorem holds.[10]

Let us briefly discuss some critical features of the symmetry problem of Eq. (2.1). As is well known, once $f_0(t)$, $f_1(t)$, and $f_2(t)$ are given, the symmetry group of such an equation is realized by a set of point transformations,

$$t'=T(t,x),\quad x'=S(t,x), \tag{2.2}$$

with nonvanishing Jacobian, endowed with the property of leaving Eq. (2.1) form invariant. Namely, the following equivalence holds:

$$L(t)x=f_0(t)\Leftrightarrow L(t')x'=f_0(t'), \tag{2.3}$$

upon the transformation of space-time variables (2.2), where $L(t)$ denotes the corresponding linear operator. If one considers the first and second extensions of these transformations (i.e., $\dot{x}\rightarrow\dot{x}'$ and $\ddot{x}\rightarrow\ddot{x}''$), substitutes them into Eq. (2.3), and separates the coefficients for the different powers of $\dot{x}$ (as usual[1,11]), one gets a system of four coupled nonlinear second-order partial differential equations for T and S. These equations (which we here omit, for the sake of briefness, cf., for instance, Ref. 9) are the starting point in the similarity analysis of *finite* point transformations of a linear second-order differential equation like (2.1). In each concrete instance, that is, when $f_0(t)$, $f_1(t)$, and $f_2(t)$ are given functions, the general solution of such nonlinear system for T and S affords a realization of a Lie group having *no more* than eight essential parameters.[12] These parameters may be adjusted by means of the integration constants and a set of admissible "initial conditions" introduced at a fixed ordinary event (t_0,x_0). In this fashion one gets a reasonable parametrization of the point transformation group. (This *finite similarity method* was used successfully in our previous paper,[9] for the equation $\ddot{x}+\omega^2x=0$.)

It is clear that for the case of the *general* linear equation (2.1) the familar similarity methods are completely useless because further analysis of the nonlinear equations for T and S (or of the linearized infinitesimal version thereof[11]) would require the knowledge of $f_0(t)$, $f_1(t)$, and $f_2(t)$. Since in this paper we are interested in the symmetry properties of the whole class of differential equations of the kind (2.1), instead of a specific member of this class, the problem must be faced *ab initio* under a different perspective, and recourse to techniques that differ from the usual tools of similarity analysis of differential equations[1,11] seems to be unavoidable in this case.

III. THE PROJECTIVE GROUP IN THE $(\hat{t},\hat{x})$ PLANE

The main ideas of our approach follow. In the theory of the second-order linear differential equation (2.1) one extracts as many properties of the solution as possible by considering some changes of the dependent (or the independent) variable, which reduce the number of essentially distinct types of equations. A well known example is the substitution

$$y(t)=x(t)\exp\left\{\frac{1}{2}\int_{t_0}^{t}dt'f_2(t')\right\} \tag{3.1}$$

that transforms Eq. (2.1) into

$$\ddot{y}+\left\{f_1(t)-\frac{1}{2}\dot{f}_2(t)-\frac{1}{4}(f_2(t))^2\right\}y$$
$$=f_0(t)\exp\left\{\frac{1}{2}\int_{t_0}^{t}dt'f_2(t')\right\}. \tag{3.2}$$

Another example is the Liouville transformation (of the independent variable), which somehow inverts the tranformation above. In the technique of changing the form of a differential equation into another, care must be exercised in verifying the one-to-one nature of the transformation of variables. However, one does not need the luxury of an explicit representation of the solution $x=w(t)$ *in terms of* $f_0(t)$, $f_1(t)$, and $f_2(t)$, wherefrom the power of this technique stems.[13] In the same spirit, in order to study the symmetry properties of Eq. (2.1) we shall begin by *reducing its form to the simplest one*.

First, let us settle our notation. Henceforth, $u_1(t)$ and $u_2(t)$ are two given linearly independent solutions of the corresponding homogeneous equation, i.e.,

$$L(t)u_j(t)=0, \tag{3.3}$$

for $j=1,2$, with

$$\dot{u}_1(t)u_2(t)-u_1(t)\dot{u}_2(t)\neq0. \tag{3.4}$$

We shall write $u_P(t)$ to denote a *particular* solution of Eq. (2.1). Hence for the complete primitive of the inhomogeneous equation (2.1), we write

$$x=w(t)=\alpha u_1(t)+\beta u_2(t)+u_P(t), \tag{3.5}$$

where α and β are two arbitrary real constants.

We next introduce the following lemmas.

Lemma I: Within those subintervals of $t_1<t<t_2$ where $u_2(t)\neq0$, one has

$$\left(\frac{u_1}{u_2}\right)^{\cdot\cdot}+\left(2\frac{\dot{u}_2}{u_2}+f_2\right)\left(\frac{u_1}{u_2}\right)^{\cdot}=0 \tag{3.6}$$

and

$$\left(\frac{w-u_P}{u_2}\right)^{\cdot\cdot}+\left(2\frac{\dot{u}_2}{u_2}+f_2\right)\left(\frac{w-u_P}{u_2}\right)^{\cdot}=0. \tag{3.7}$$

Here we have written $(f)^{\cdot}$, instead of $\dot{f}$, for convenience. The reader can prove Lemma I rather easily.

Lemma II: The definition of new variables,

$$\hat{t}=\tau(t):=u_1(t)/u_2(t),\quad \hat{x}=[x-u_P(t)]/u_2(t) \tag{3.8}$$

(within its domain of validity), entails a local transformation of space-time coordinates

Indeed, the Jacobian of Eq. (3.8) is given by

$$\frac{\partial(\hat{t},\hat{x})}{\partial(t,x)}=\frac{\dot{u}_1(t)u_2(t)-u_1(t)\dot{u}_2(t)}{(u_2(t))^3}\neq 0, \tag{3.9}$$

and the corresponding inverse transformation is of the general form

$$t=\tau^{-1}(\hat{t}),\quad x=A(\hat{t})\hat{x}+B(\hat{t}), \tag{3.10}$$

where, clearly,

$$\dot{\tau}^{-1}(\hat{t}):=\frac{dt}{d\hat{t}}=\left(\left(\frac{u_1}{u_2}\right)^{\cdot}\right)^{-1},$$
$$A(\hat{t}):=u_2(\tau^{-1}(\hat{t})),\quad B(\hat{t}):=u_P(\tau^{-1}(\hat{t})). \tag{3.11}$$

Next, we take advantage of these rather simple facts, and we assert the following theorems.

Theorem I: The local transformation of space-time coordinates, stated in Eq. (3.8), transforms the equation of motion

$$\ddot{x}+f_2(t)\dot{x}+f_1(t)x-f_0(t)=0 \tag{3.12}$$

into

$$\ddot{\hat{x}}=0, \tag{3.13}$$

where $\ddot{\hat{x}}=d\dot{\hat{x}}/d\hat{t}$ and $\dot{\hat{x}}=d\hat{x}/d\hat{t}$.

Proof: In fact, the first and second extensions of the transformation (3.8) are given by

$$\dot{\hat{x}}=\left(\left(\frac{u_1}{u_2}\right)^{\cdot}\right)^{-1}\left(\frac{x-u_P}{u_2}\right)^{\cdot}, \tag{3.14}$$

$$\ddot{\hat{x}}=\left(\left(\frac{u_1}{u_2}\right)^{\cdot}\right)^{-2}\left(\frac{x-u_P}{u_2}\right)^{\cdot\cdot}-\left(\left(\frac{u_1}{u_2}\right)^{\cdot}\right)^{-3}\left(\frac{u_1}{u_2}\right)^{\cdot\cdot}\left(\frac{x-u_P}{u_2}\right)^{\cdot}, \tag{3.15}$$

respectively. Therefore, in particular, "on the world lines" $x=w(t)$ as a consequence of Lemma I, Eq. (3.15) yields (3.13). This ends the proof.

Theorem II: The inverse local transformation of space-time coordinates, defined in Eqs. (3.10) and (3.11), changes the differential equation (3.13) into (3.12).

The proof of this theorem is straightforward (albeit rather lengthy). We shall briefly refer to transformation (3.8) as a $\mathscr{F}$ *transformation.* Of course, since both solutions, $u_1(t)$ and $u_2(t)$, are on the same footing, it is clear that similar results hold for a (u_2/u_1) transformation scheme in those subintervals of $t_1<t<t_2$ where $u_1(t)\neq 0$, *mutatis mutandi.* Hence one may cover any continuous space-time band with a set of overlapping $\mathscr{F}$ transformations producing a set of overlapping $(\hat{t},\hat{x})$ coordinate patches, as the case may be.

These are far-reaching results, for it is immediate that they bring to the fore the *projective group of the plane* as playing an outstanding role in the description of the point symmetry properties of a linear second-order differential equation like (2.1). Let us discuss this subject rather briefly. Having performed the local $\mathscr{F}$ transformation, if one next performs a projective transformation, $\mathscr{P}_2(q)$: $(\hat{t},\hat{x})\overset{q}{\rightarrow}(\hat{t}\,',\hat{x}')$, say,

$$\hat{t}\,'=\frac{q^1\hat{t}+q^2\hat{x}+q^3}{q^7\hat{t}+q^8\hat{x}+1},\quad \hat{x}'=\frac{q^4\hat{t}+q^5\hat{x}+q^6}{q^7\hat{t}+q^8\hat{x}+1}, \tag{3.16}$$

one certainly obtains

$$\ddot{\hat{x}}=0\Rightarrow\ddot{\hat{x}}'=0. \tag{3.17}$$

Hence if one finally performs the inverse transformation, $\mathscr{F}^{-1}$: $(\hat{t}\,',\hat{x}')\rightarrow(t',x')$, as in Eqs. (3.10) and (3.11), i.e.,

$$t'=\tau^{-1}(\hat{t}\,'),\quad x'=A(\hat{t}\,')\hat{x}'+B(\hat{t}\,'), \tag{3.18}$$

[with $(\hat{t}\,',\hat{x}')$ as given in Eq. (3.16)] according to Theorem II, one certainly gets

$$\ddot{\hat{x}}'=0\Rightarrow\ddot{x}'+f_2(t')\dot{x}'+f_1(t')x'-f_0(t')=0. \tag{3.19}$$

Thus the transformation

$$\mathscr{F}^{-1}\mathscr{P}_2(q)\mathscr{F}:\quad (t,x)\overset{q}{\rightarrow}(t',x') \tag{3.20}$$

gives us the space-time realization of the full point symmetry group of a one-dimensional linear Newtonian system. Since the $\mathscr{F}$ transformations do *not* form a Lie group, and since, clearly,

$$\begin{aligned}(\mathscr{F}^{-1}\mathscr{P}_2(q')\mathscr{F})(\mathscr{F}^{-1}\mathscr{P}_2(q)\mathscr{F})&\\ =\mathscr{F}^{-1}(\mathscr{P}_2(q')\mathscr{P}_2(q))\mathscr{F}&\\ =\mathscr{F}^{-1}\mathscr{P}_2(g(q';q))\mathscr{F},&\end{aligned} \tag{3.21}$$

where $g^a(q';q)=q''^a$, $a=1,...,8$, denotes the binary composition law of the parameters *of the projective group*, we have shown the following theorem.

Theorem III: The full point symmetry group of Eq. (2.1) corresponds to those space-time realizations of the projective group in plane $(\mathscr{P}_2)$ which are of the form $\mathscr{F}^{-1}\mathscr{P}_2\mathscr{F}$.

To end this section, let us epitomize and exhibit the eight-parameter transformations $(t,x)\overset{q}{\rightarrow}(t',x')$ that keep invariant the equation of motion of a one-dimensional linear system. After substitution from Eqs. (3.8) into (3.16), we obtain

$$\begin{aligned}\hat{t}\,'&=\frac{u_1(t')}{u_2(t')}=\frac{q^1u_1(t)+q^3u_2(t)+q^2(x-u_P(t))}{q^7u_1(t)+q^8(x-u_P(t))+u_2(t)},\\ \hat{x}'&=\frac{x'-u_P(t')}{u_2(t')}=\frac{q^4u_1(t)+q^6u_2(t)+q^5(x-u_P(t))}{q^7u_1(t)+q^8(x-u_P(t))+u_2(t)}.\end{aligned} \tag{3.22}$$

In this way, using Eqs. (3.18) [cf. also Eqs. (3.11)], one gets the final answer,

$$\begin{aligned}t'&=\tau^{-1}\left(\frac{q^1u_1(t)+q^3u_2(t)+q^2(x-u_P(t))}{q^7u_1(t)+q^8(x-u_P(t))+u_2(t)}\right)\\ x'&=\frac{q^4u_1(t)+q^6u_2(t)+q^5(x-u_P(t))}{q^7u_1(t)+q^8(x-u_P(t))+u_2(t)}\,u_2\\ &\quad\times\left(\tau^{-1}\left(\frac{q^1u_1(t)+q^3u_2(t)+q^2(x-u_P(t))}{q^7u_1(t)+q^8(x-u_P(t))+u_2(t)}\right)\right)\end{aligned} \tag{3.23}$$

$$+\,u_P\left(\tau^{-1}\left(\frac{q^1u_1(t)+q^3u_2(t)+q^2(x-u_P(t))}{q^7u_1(t)+q^8(x-u_P(t))+u_2(t)}\right)\right).$$

One can easily check the identity in Eqs. (3.23) against the identity in Eqs. (3.16); i.e., one sets $q^1=q^5=1$, and $q^2=q^3=q^4=q^6=q^7=q^8=0$. Furthermore, one observes that using a (u_2/u_1) scheme for the $\mathscr{F}$ transformations, instead of the (u_1/u_2) scheme used in this paper, is tantamount to a trivial reparametrization of the projective group (as it should be).

It must be borne in mind that, according to Eq. (3.21), the transformations of space-time coordinates stated in Eqs. (3.23) carry a *faithful* realization of the eight-parameter group $\mathscr{P}_2$, within their domain of definition. Hence we have shown the following theorem.

Theorem IV: The full point symmetry group of Eq. (2.1) is isomorphic with the projective group $\mathscr{P}_2$.

IV. SL(3,R)

We are in position to discuss the issue of SL$(3,R)$ in connection with the symmetry properties of the general linear equation of motion (2.1). To this end, we follow the same arguments used in Ref. 9, which rest exclusively in the linear character of the differential equation.

Let us handle the transformations (3.22) in a "compact" fashion. Therefore we write

$$\begin{pmatrix} u_1(t')/u_2(t') \\ [x'-u_P(t')]/u_2(t') \\ 1 \end{pmatrix} = \phi\begin{pmatrix} q^1 & q^2 & q^3 \\ q^4 & q^5 & q^6 \\ q^7 & q^8 & 1 \end{pmatrix}\begin{pmatrix} u_1(t)/u_2(t) \\ [x-u_P(t)]/u_2(t) \\ 1 \end{pmatrix}, \tag{4.1}$$

where

$$\phi=\left(1+q^7\frac{u_1(t)}{u_2(t)}+q^8\frac{x-u_P(t)}{u_2(t)}\right)^{-1}. \tag{4.2}$$

We then rewrite Eq. (4.1) symbolically, to read

$$\mathbf{v}'=\phi(\mathbf{v};q)\mathbf{M}(q)\cdot\mathbf{v}, \tag{4.3}$$

which meaning is clear. Let us observe that this last equation may be written also in the form

$$\mathbf{v}'=\mathbf{M}(q)\cdot\mathbf{v}/(\mathbf{M}(q)\cdot\mathbf{v})_3\,, \tag{4.4}$$

where $(\mathbf{M}(q)\cdot\mathbf{v})_3$ stands for the third row in $\mathbf{M}(q)\cdot\mathbf{v}$. Equation (4.4) shows neatly the projective nature of the transformation (4.1).

Next, we need the Jacobian $J=\partial(t',x')/\partial(t,x)$ of Eqs. (3.23). A straightforward calculation yields

$$J=\frac{u_2(t')(u_1(t)/u_2(t))^{\cdot}}{u_2(t)(u_1(t')/u_2(t'))^{\cdot}}\,\phi^3\det(\mathbf{M}(q))\,. \tag{4.5}$$

In consequence, upon performing two successive transformations, we require that

$$\mathbf{v}''=\phi''\mathbf{M}''\cdot\mathbf{v}=\phi'\phi\mathbf{M}'\cdot\mathbf{M}\cdot\mathbf{v} \tag{4.6}$$

holds for *every* column $\mathbf{v}=(u_1/u_2,(x-u_P)/u_2,1)$ (transposed), as this is a necessary consequence of the group property of these transformations. Of course, in Eq. (4.6) we have written $\phi'=\phi(\mathbf{v}';q')$, $\phi''=\phi(\mathbf{v};q'')$, $\mathbf{M}'=\mathbf{M}(q')$, and $\mathbf{M}''=\mathbf{M}(q'')$, where, clearly, $q''^a=g^a(q';q)$, $a=1,\ldots,8$, as one obtains from the projective group [cf. Eq. (3.22)]. Since $J''=J'J$, Eq. (4.5) yields immediately

$$\phi''^3\det(\mathbf{M}'')=(\phi'\phi)^3\det(\mathbf{M}')\det(\mathbf{M}); \tag{4.7}$$

i.e., the expression

$$(\phi''/\phi'\phi)^3=\det(\mathbf{M}'\cdot\mathbf{M})/\det(\mathbf{M}'') \tag{4.8}$$

is a function of q' and q only, for it does *not* depend on the $\mathbf{v}$'s. Hence from Eq. (4.6) one readily gets

$$\frac{\mathbf{M}''}{(\det(\mathbf{M}''))^{1/3}}=\frac{\mathbf{M}'}{(\det(\mathbf{M}'))^{1/3}}\cdot\frac{\mathbf{M}}{(\det(\mathbf{M}))^{1/3}}, \tag{4.9}$$

or, more explicitly,

$$\frac{\mathbf{M}(g(q';q))}{(\det\{\mathbf{M}(g(q';q))\})^{1/3}} =\frac{\mathbf{M}(q')}{(\det\{\mathbf{M}(q')\})^{1/2}}\cdot\frac{\mathbf{M}(q)}{(\det\{\mathbf{M}(q)\})^{1/3}}, \tag{4.10}$$

where, obviously, these *eight-parameter* matrices are elements of SL$(3,R)$,

$$\frac{\mathbf{M}(q)}{(\det\{\mathbf{M}(q)\})^{1/3}}\in\mathrm{SL}(3,R)\,. \tag{4.11}$$

This result exhibits the fact that every element in $\mathscr{F}^{-1}\mathscr{P}_2\mathscr{F}$ corresponds to an element in SL$(3,R)$, and that the group law for the binary combination of the elements in $\mathscr{F}^{-1}\mathscr{P}_2\mathscr{F}$ is the same as in SL$(3,R)$ [i.e., $g(q';q)=q''$].

This is a rather natural result indeed, since the groups $\mathscr{P}_2$ and SL$(3,R)$ are *isomorphic*: $\mathscr{P}_2$ is a realization of SL$(3,R)$ and SL$(3,R)$ is a representation of $\mathscr{P}_2$. In fact, if one writes the generic element of the projective transformation (3.16) in the form of Eq. (4.1), namely,

$$\begin{pmatrix} \hat{t}' \\ \hat{x}' \\ 1 \end{pmatrix}=(q^7\hat{t}+q^8\hat{x}+1)^{-1}\begin{pmatrix} q^1 & q^2 & q^3 \\ q^4 & q^5 & q^6 \\ q^7 & q^8 & 1 \end{pmatrix}\begin{pmatrix} \hat{t} \\ \hat{x} \\ 1 \end{pmatrix}, \tag{4.12}$$

since the Jacobian of (3.16) is given by

$$\frac{\partial(\hat{t}',\hat{x}')}{\partial(\hat{t},\hat{x})}=(q^7\hat{t}+q^8\hat{x}+1)^{-3}\det(\mathbf{M}(q))\,, \tag{4.13}$$

it is clear that, if one follows the same argumentation used in connection with Eqs. (4.1) and (4.5), one will equally arrive at Eq. (4.10) [as it must be, for Eq. (4.12) is a special instance of Eq. (4.1)]. Hence to every element of $\mathscr{P}_2$ there corresponds one (and only one) well defined 3×3 unimodular matrix [cf. Eq. (4.11)]. Moreover, the group law of $\mathscr{P}_2$ is the same as the group law of these image matrices. Thus $\mathscr{P}_2$ is a subgroup of SL$(3,R)$. *Conversely*, it can be shown (though the proof is much more involved) that every 3×3 unimodular real matrix yields a unique *triplet of equivalent projective transformations* (TEPT) and that this mapping preserves the group law. Moreover, the ordered TEPT's constitute a group that is isomorphic with $\mathscr{P}_2$. (See the Appendix for a sketchy proof of these features.) Thus SL$(3,R)$ is a subgroup of $\mathscr{P}_2$. Hence both groups are isomorphic, and SL$(3,R)$ is the full point symmetry group of all conceivable one-dimensional linear Newtonian systems.

V. SOME MISCELLANEOUS EXAMPLES

Interestingly enough, from the standpoints of mechanics and applied mathematics, Eqs. (3.23) afford an effective tool for the *construction* of the full point symmetry transformations of any given linear ordinary second-order differential equation. In this section we present some interesting instances, taken from elementary mechanics and analysis.

(a) *Free particle.* This is a trivial check of the formalism. We have $\ddot{x}=0$; i.e., we take $u_1(t)=t$, $u_2(t)=1$, while, clearly $u_P(t)=0$. Hence Eqs. (3.22) and (3.23) yield immediately the projective transformation, as it must be

(b) *Free falling particle.* Now we set $\ddot{x}+g=0$. Thus we have, for instance, $u_1(t)=t$, $u_2(t)=1$, and $u_P(t)=-\frac{1}{2}gt^2$, wherefrom we easily obtain the transformation

$$t'=\frac{q^1t+q^2(x+\frac{1}{2}gt^2)+q^3}{q^7+q^8(x+\frac{1}{2}gt^2)+1},$$

$$x'=\frac{q^4t+q^5(x+\frac{1}{2}gt^2)+q^6}{q^7t+q^8(x+\frac{1}{2}gt^2)+1} - \frac{1}{2}g\left(\frac{q^1t+q^2(x+\frac{1}{2}gt^2)+q^3}{q^7t+q^8(x+\frac{1}{2}gt^2)+1}\right)^2. \tag{5.1}$$

The reader can check Eq. (5.1) against the fundamental requirement: $\ddot{x}+g=0\Leftrightarrow\ddot{x}'+g=0$. Let us observe that any other admissible choice for the u's corresponds merely to a reparametrization of the projective group. This is a *general rule* of this approach.

(c) *Simple harmonic oscillator.* We quickly obtain the space-time point symmetries of $\ddot{x}+\omega^2x=0$. Let us take $u_1(t)=\sin\omega t$, $u_2(t)=\cos\omega t$, $u_P(t)=0$. Then, from Eqs. (3.22) and (3.23), after some elementary manipulations, one gets

$$t'=\frac{1}{\omega}\arctan\left(\frac{q^1\sin\omega t+q^2x+q^3\cos\omega t}{q^7\sin\omega t+q^8x+\cos\omega t}\right),\quad x'=\frac{q^4\sin\omega t+q^5x+q^6\cos\omega t}{((q^1\sin\omega t+q^2x+q^3\cos\omega t)^2+(q^7\sin\omega t+q^8x+\cos\omega t)^2)^{1/2}}. \tag{5.2}$$

This transformation of variables is equivalent to the transformation obtained in Paper I, within a suitable reparametrization of the group. However, the connection of the formulas presented in that paper with the projective group is not as transparent as in Eqs. (5.2).

(d) *A forced harmonic oscillator.* Let us consider the inhomogeneous equation of motion

$$\ddot{x}+\omega^2x=f_0\sin\Omega t, \tag{5.3}$$

where f_0 is a constant. We take $u_1(t)=\sin\omega t$, $u_2(t)=\cos\omega t$, as before, and use

$$u_P(t)=-[f_0/(\Omega^2-\omega^2)]\sin\Omega t. \tag{5.4}$$

Hence from Eqs. (3.22) and (3.23) one obtains the following rather formidable transformation of variables:

$$\begin{aligned} t'&=\frac{1}{\omega}\arctan\left(\frac{q^1\sin\omega t+q^2(x-u_P(t))+q^3\cos\omega t}{q^7\sin\omega t+q^8(x-u_P(t))+\cos\omega t}\right),\\ x'&=\frac{q^4\sin\omega t+q^5(x-u_P(t))+q^6\cos\omega t}{((q^1\sin\omega t+q^2(x-u_P(t))+q^3\cos\omega t)^2+(q^7\sin\omega t+q^8(x-u_P(t))+\cos\omega t)^2)^{1/2}}\\ &\quad-\frac{f_0}{\Omega^2-\omega^2}\sin\left(\frac{\Omega}{\omega}\arctan\left(\frac{q^1\sin\omega t+q^2(x-u_P(t))+q^3\cos\omega t}{q^7\sin\omega t+q^8(x-u_P(t))+\cos\omega t}\right)\right), \end{aligned} \tag{5.5}$$

which leaves invariant Eq. (5.3). Of course, it is very tedious to check this fact with the eight parameters included. However, the reader can test Eqs. (5.5) against the symmetries of (5.3) by considering at least the transformations that belong to the one-parameter subgroups.

(e) *Damped harmonic oscillator.* We consider the equation of motion $\ddot{x}+2\lambda\dot{x}+\omega^2x=0$, with $u_1(t)=e^{-\lambda t}\sin\Omega t$, $u_2(t)=e^{-\lambda t}\cos\Omega t$, where $\Omega=\sqrt{\omega^2-\lambda^2}$, and $u_P(t)=0$. Hence after a straightforward calculation we get the point symmetry transformations of this equation; i.e.,

$$\begin{aligned} t'&=\frac{1}{\Omega}\arctan\left(\frac{q^1\sin\Omega t+q^2xe^{\lambda t}+q^3\cos\Omega t}{q^7\sin\Omega t+q^8xe^{\lambda t}+\cos\Omega t}\right),\\ x'&=\frac{q^4\sin\Omega t+q^5xe^{\lambda t}+q^6\cos\Omega t}{((q^1\sin\Omega t+q^2xe^{\lambda t}+q^3\cos\Omega t)^2+(q^7\sin\Omega t+q^8xe^{\lambda t}+\cos\Omega t)^2)^{1/2}}\\ &\quad\times\exp\left(-\frac{\lambda}{\Omega}\arctan\left(\frac{q^1\sin\Omega t+q^2xe^{\lambda t}+q^3\cos\Omega t}{q^7\sin\Omega t+q^8xe^{\lambda t}+\cos\Omega t}\right)\right). \end{aligned} \tag{5.6}$$

(f) *Falling particle in a viscous media.* Now, let the equation be $\ddot{x}+\lambda\dot{x}=-g$. We take $u_1(t)=e^{-\lambda t}$, $u_2(t)=1$, $u_P(t)=-(g/\lambda)t$. Therefore we obtain the following symmetry transformation of variables for this equation of motion:

$$t' = -\frac{1}{\lambda}\ln\left(\frac{q^1e^{-\lambda t}+q^2(x+(g/\lambda)t)+q^3}{q^7e^{-\lambda t}+q^8(x+(g/\lambda)t)+1}\right),$$
$$x' = \frac{q^4e^{-\lambda t}+q^5(x+(g/\lambda)t)+q^6}{q^7e^{-\lambda t}+q^8(x+(g/\lambda)t)+1}+\frac{g}{\lambda^2}\ln\left(\frac{q^1e^{-\lambda t}+q^2(x+(g/\lambda)t)+q^3}{q^7e^{-\lambda t}+q^8(x+(g/\lambda)t)+1}\right). \tag{5.7}$$

(g) *Symmetries of* $\ddot{x}+t^{-1}\dot{x}-t^{-2}x=0$. Finally, we also consider one example of a linear differential equation with time-dependent coefficients. Although not very interesting from the point of view of mechanics, for the sake of simplicity let us consider the following:

$$\ddot{x}+(1/t)\dot{x}-(1/t^2)x=0\,, \tag{5.8}$$

which has a regular singular point at $t=0$. We set $u_1(t)=t$, $u_2(t)=t^{-1}$, and $u_P(t)=0$. Thus one readily obtains

$$t' = \left(\frac{q^1t^2+q^2tx+q^3}{q^7t^2+q^8tx+1}\right)^{1/2},$$
$$x' = \frac{q^4t^2+q^5tx+q^6}{((q^7t^2+q^8tx+1)(q^1t^2+q^2tx+q^3))^{1/2}}, \tag{5.9}$$

for the realization of the full point symmetry group of Eq. (5.8).

In all these examples one can find very easily the corresponding realizations of the Lie algebra of SL(3,R) by considering the monoparametric transformations of variables to the first order of approximation in the parameter one handles. In a forthcoming paper we shall discuss the general Lie algebra of the symmetry group of Eq. (2.1) by means of the formalism introduced in the present paper.

Work is in progress concerning SL(3,R) quantum kinematics and the geometric quantization of linear Newtonian systems in two-dimensional space-time, according to the general results obtained in this paper.[14]

AKCNOWLEDGMENTS

This work was supported in part by DGI(UCV), Grant 123.752/86, in part by DIUC(PUC), Grant 50/87, and in part by FNDCYT, Contract 1106/85.

APPENDIX

Here we append the proof that SL(3,R) is a faithful representation of $\mathscr{P}_2$ and, hence, that both groups are isomorphic. For the sake of briefness, we shall discuss this subject in a rather sketchy way.

One has the following mapping of GL(3,R) onto SL(3,R) [cf. Eq. (4.11)]:

$$\mathrm{M}\in\mathrm{GL}(3,R) \Rightarrow \mathrm{M}/(\det(\mathrm{M}))^{1/3}\in\mathrm{SL}(3,R)\,. \tag{A1}$$

Hence one shows rather easily that a necessary and sufficient condition for two elements of GL(3,R) to yield the same element of SL(3,R) is that they be linearly dependent.

On the other hand, it is clear that, in general, every element M∈GL(3,R) affords three projective transformations in two-dimensional projective space, say,

$$\frac{x'}{z'} = \frac{M_{33}^{-1}(M_{11}(x/z)+M_{12}(y/z)+M_{13})}{M_{33}^{-1}(M_{31}(x/z)+M_{32}(y/z))+1},$$
$$\frac{y'}{z'} = \frac{M_{33}^{-1}(M_{21}(x/z)+M_{22}(y/z)+M_{23})}{M_{33}^{-1}(M_{31}(x/z)+M_{32}(y/z))+1}, \tag{A2}$$

$$\frac{z'}{y'} = \frac{M_{22}^{-1}(M_{33}(z/y)+M_{31}(x/y)+M_{32})}{M_{22}^{-1}(M_{23}(z/y)+M_{21}(x/y))+1},$$
$$\frac{x'}{y'} = \frac{M_{22}^{-1}(M_{13}(z/y)+M_{11}(x/y)+M_{12})}{M_{22}^{-1}(M_{23}(z/y)+M_{21}(x/y))+1}, \tag{A3}$$

$$\frac{y'}{x'} = \frac{M_{11}^{-1}(M_{22}(y/x)+M_{23}(z/x)+M_{21})}{M_{11}^{-1}(M_{12}(y/x)+M_{13}(z/x))+1},$$
$$\frac{z'}{x'} = \frac{M_{11}^{-1}(M_{32}(y/x)+M_{33}(z/x)+M_{31})}{M_{11}^{-1}(M_{12}(y/x)+M_{13}(z/x))+1}, \tag{A4}$$

where $x'_j = M_{jk}\,x_k$. [The special cases, when $M_{11}=0$, $M_{22}=0$, or $M_{33}=0$, correspond to holonomic constraints which reduce the number of parameters of GL(3,R); thus one can disregard these loci without loss of generality.] Furthermore, it is immediate that the rule of combination of these induced projective transformations preserves the group law of GL(3,R). Of course, the three projective transformations induced by M∈GL(3,R) are *not* independent. (This feature will be discussed presently.) Now it can be shown that the necessary and sufficient condition for two elements of GL(3,R) to induce the same triplet of projective transformations is that they are linearly dependent. Therefore, the important conclusion follows: *every element of* SL(3,R) *induces one, and only one, triplet of projective transformations.*

We next consider the following mapping of the two-dimensional projective space $\{(\alpha,\beta)\}$ onto itself:

$$(\alpha,\beta)^* = (1/\beta,\alpha/\beta), \tag{A5}$$

i.e.,

$$(\alpha,\beta)^{**} = (1/\beta,\alpha/\beta)^* = (\beta/\alpha,1/\alpha), \tag{A6}$$

$$(\alpha,\beta)^{***} = (\beta/\alpha,1/\alpha)^* = (\alpha,\beta). \tag{A7}$$

Obviously, the sets $\{(\alpha,\beta)\}$, $\{(\alpha,\beta)^*\}$, and $\{(\alpha,\beta)^{**}\}$ are *equivalent* systems of homogeneous coordinates in projective space [i.e., Eq. (A5) is just a transformation of coordinates]. Then, let $P(p)$ denote a projective transformation, with parameters $p=(p_1,\ldots,p_8)$; namely,

$$P(p):\quad \alpha' = \frac{p_1\alpha+p_2\beta+p_3}{p_7\alpha+p_8\beta+1},$$
$$\beta' = \frac{p_4\alpha+p_5\beta+p_6}{p_7\alpha+p_8\beta+1}. \tag{A8}$$

We shall say that three projective transformations, $P(p)(\alpha,\beta)$, $P(p^*)(\alpha,\beta)^*$, and $P(p^{**})(\alpha,\beta)^{**}$, form a *triplet of equivalent projective transformations* (TEPT) if they preserve the (*) mapping for the transformed coordinates.

Of course, the three elements in a TEPT are not independent. Indeed, a straightforward calculation yields

$$p_1^* = p_5^{-1},\quad p_2^* = p_7 p_5^{-1},\quad p_3^* = p_8 p_5^{-1},\quad p_4^* = p_3 p_5^{-1},$$
$$p_5^* = p_1 p_5^{-1},\quad p_6^* = p_2 p_5^{-1},\quad p_7^* = p_6 p_5^{-1},\quad p_8^* = p_4 p_5^{-1}, \tag{A9}$$

and

$$\begin{aligned} &p_1^{**} = p_5 p_1^{-1},\quad p_2^{**} = p_6 p_1^{-1},\quad p_3^{**} = p_4 p_1^{-1},\\ &p_4^{**} = p_8 p_1^{-1},\quad p_5^{**} = p_1^{-1},\quad p_6^{**} = p_7 p_1^{-1},\\ &p_7^{**} = p_2 p_1^{-1},\quad p_8^{**} = p_3 p_1^{-1}, \end{aligned} \tag{A10}$$

for the parameters of $P^* = P(p^*)$ and $P^{**} = P(p^{**})$, respectively. One observes that $p^{**} = (p^*)^*$ and $p^{***} = p$, as it should be; i.e., it follows that

$$P^{**} = (P^*)^*, \tag{A11}$$

$$P^{***} = P. \tag{A12}$$

Also

$$P^* = P \Leftrightarrow p_1 = p_5 = 1,\quad p_2 = p_6 = p_7,\quad p_3 = p_4 = p_8. \tag{A13}$$

Moreover, one has

$$(P'P)^* = P'^* P^*, \tag{A14}$$

$$(P^{-1})^* = (P^*)^{-1}, \tag{A15}$$

$$I = I^* = I^{**} \tag{A16}$$

(where I stands for the identity in $\mathscr{P}_2$), and

$$P'^* = P^* \Leftrightarrow P' = P. \tag{A17}$$

Therefore the mapping $P \rightarrow P^*$ is an automorphism of $\mathscr{P}_2$.

Every element of $\mathscr{P}_2$ belongs to one TEPT. However, all three members of a TEPT appear on an equal footing, so that any member of the TEPT can be taken as the representative that labels the triplet and from which the TEPT can be constructed. Thus one can take the first member to this end, and write

$$T(P) = (P, P^*, P^{**}), \tag{A18}$$

while considering the TEPT's as an *ordered triplets*. In fact, it can be shown that in this way (and only in this way) the following TEPT product:

$$T(P')T(P) = T(P'P) \tag{A19}$$

is unambiguous. So one establishes an *isomorphism* between the group of *ordered* TEPT's and $\mathscr{P}_2$. (The set of *unordered* TEPT's do *not* constitute a group, because their product would be ambiguous.) This argument becomes strengthened if one observes that one has a homomorphism of $\mathscr{P}_2$ onto {TEPT} whose kernel is $\{I\}$ [cf. Eq. (A16)].

Finally, a glance at Eqs. (A2)–(A4) shows that the three projective transformations associated with an element of GL$(3,R)$ constitute a TEPT. Hence, every element of SL$(3,R)$ corresponds to one, and only one, TEPT; and therefore, according to the previous argument, SL$(3,R)$ carries a faithful representation of $\mathscr{P}_2$. Namely, both groups are isomorphic. This ends the proof.

[1] L. V. Ovsiannikov, *Group Analysis of Differential Equations* (Academic, New York, 1982).

[2] A brief survey of recent work can be seen in P. Rudra, Pramana **23**, 445 (1984).

[3] A lucid discussion of this subject can be found in G. E. Prince, Bull. Austral. Math. Soc. **25**, 309 (1982); **27**, 53 (1983); J. Phys. A: Math. Gen. **16**, L105 (1983); Bull. Austral. Math. Soc. **32**, 299 (1985).

[4] J. F. Cariñena and L. A. Ibort, J. Phys. A: Math. Gen. **16**, 1 (1983).

[5] Compare, for instance, G. H. Katzin and J. Levine, J. Math. Phys. **9**, 8 (1968); G. S. Hall, Lett. Nuovo Cimento **6**, 46 (1973); G. H. Katzin, J. Math. Phys. **14**, 1213 (1973); cf., also, Refs. 2 and 3, and work quoted therein. The recent literature on the issue of symmetries of differential equations of motion is so extended that our references do not pretend to be complete.

[6] G. E. Prince and C. J. Eliezer, J. Phys. A: Math. Gen. **14**, 587 (1981).

[7] C. E. Wulfman and B. G. Wybourne, J. Phys. A: Math. Gen. **9**, 507 (1976).

[8] Later on it was found that SL$(3,R)$ is also the complete symmetry group of the one-dimensional *time-dependent* harmonic oscillator; cf. P. G. L. Leach, J. Math. Phys. **21**, 300 (1980). Furthermore, it has been shown that the full symmetry group of the time-dependent n-dimensional harmonic oscillator is SL$(n+2,R)$; cf. G. E. Prince and C. J. Eliezer, J. Phys. A: Math. Gen. **13**, 815 (1980).

[9] M. Aguirre and J. Krause, J. Phys. A: Math. Gen. **20**, 3553 (1987).

[10] Compare, E. L. Ince, *Ordinary Differential Equations* (Dover, New York, 1956).

[11] See. G. W. Bluman and J. D. Cole, *Similarity Methods for Differential Equations* (Springer, New York, 1974). Compare, also, M. Aguirre and J. Krause, J. Math. Phys. **25**, 210 (1984); **26**, 593 (1985).

[12] Compare, for instance, L. Bianchi, *Lezioni sulla teoria dei gruppi finiti di transformazioni* (Zanichelli, Bologna, 1928), Chap. X.

[13] See, for instance, R. Bellman, *Stability Theory of Differential Equations* (McGraw-Hill, New York, 1953).

[14] J. Krause, J. Phys. A: Math. Gen. **18**, 1309 (1985); J. Krause, J. Math. Phys. **27**, 2922 (1986). Compare, also, J. Krause, "Galilean quantum kinematics," to appear in J. Math. Phys.

Supersymmetry, parastatistics, and operator realizations of a Lie algebra

S. N. Biswas and S. K. Soni
Department of Physics and Astrophysics, University of Delhi, Delhi-110007, India

(Received 5 May 1987; accepted for publication 9 September 1987)

The algebraic structure of parastatistics has been generalized and it is found to be consistent with supersymmetric quantum mechanics with supercharges constructed out of the generalized para-Bose and para-Fermi operators. It is further shown that the operator algebra of generalized parastatistics offers a realization of the (graded) orthosymplectic group similar to that of orthogonal and symplectic groups using conventional parastatistics.

I. INTRODUCTION

It is well known that supersymmetry expresses an invariance of the Lagrangian under simultaneous transformation of bosons and fermions, when the bosons and fermions obey canonical Bose and Fermi quantization rules. In supersymmetric quantum mechanics[1] if H is the Hamiltonian of the system and Q and $Q^\dagger$ are the supercharge and its Hermitian conjugate, then the supersymmetry generates the well-known graded algebra, namely,

$$\{Q,Q^\dagger\} = 2H, \quad [H,Q] = \{Q,Q\} = 0. \tag{1.1}$$

In the present paper we would like to address ourselves to the question of whether the supersymmetric quantum mechanics described by the graded Lie algebra (1.1) is consistent with the situation where the fermions and bosons obey the quantization rules according to parastatistics[2] instead of canonical Fermi and Bose statistics. In Sec. II we discuss this and show that the consistency of (1.1) with parastatistics requires additional algebraic relations between para-Fermi and para-Bose generators not present in the standard Green's result, but nevertheless quite consistent with canonical Fermi and Bose quantum conditions. We write down the complete algebraic structure leading to generalized parastatistics. To find the additional significance of these new relations we have discussed in the remaining sections the operator realizations of the classical graded Lie algebra based on the operators obeying generalized parastatistics. Sections III–V have been added for the sake of completeness. In Sec. III we review the constructions of the unitary algebras based on conventional Fermi and Bose operators and mention the salient points of parastatistics. In this connection, we may mention the work of Bracken and Green[3] who construct all the SU(3) multiplets with para-Fermi field operators of order 3. In Secs. IV and V we make use of the complete set of fundamental relations of parastatistics and show their consistency with operator realizations of orthogonal and symplectic algebras (which contain unitary algebras as their subalgebras). Long ago, Ryan and Sudarshan[4] had established a very direct connection between parafermion and paraboson algebras with those of the Lie algebras of orthogonal and symplectic groups, respectively. The special case of O(3) was discussed in detail by Jordan *et al.*[5] Finally in Sec. VI we generalize our observations to (graded unitary and orthosymplectic types of) classical graded Lie algebras by showing that their operator realizations are entirely consistent with generalized parastatistics.

II. THE FUNDAMENTAL RELATIONS OF GENERALIZED PARASTATISTICS

In this section we shall give the algebraic structure of generalized parastatistics. We shall motivate this by observing the consistency between this structure and that of supersymmetric quantum mechanics. The fundamental relations of generalized parastatistics are at one glance

$$[b_\alpha,[b^\dagger_\beta,b_\gamma]] = 2\delta_{\alpha\beta} b_\gamma, \tag{2.1a}$$

$$[a_i,\{a^\dagger_j,a_k\}] = 2\delta_{ij} a_k, \tag{2.1b}$$

$$[b_\alpha,\{a^\dagger_i,a_j\}] = [a_i,[b^\dagger_\alpha,b_\beta]] = 0, \tag{2.1c}$$

$$[a_i,\{b^\dagger_\alpha,a_j\}] = \{b_\alpha,\{a^\dagger_i,b_\beta\}\} = 0, \tag{2.1d}$$

$$[a_i,\{b_\alpha,a^\dagger_j\}] = 2\delta_{ij} b_\alpha, \quad \{b_\alpha,\{b^\dagger_\beta,a_i\}\} = 2\delta_{\alpha\beta} a_i. \tag{2.1e}$$

$$[b^\dagger_\alpha,[b_\beta,b_\gamma]] = 2\delta_{\alpha\beta} b_\gamma - 2\delta_{\alpha\gamma} b_\beta, \tag{2.2a}$$

$$[a^\dagger_i,\{a_j,a_k\}] = -2\delta_{ij} a_k - 2\delta_{ik} a_j, \tag{2.2b}$$

$$[a^\dagger_i,\{a_j,b_\alpha\}] = -2\delta_{ij} b_\alpha, \quad \{b^\dagger_\alpha,\{a_i,b_\beta\}\} = 2\delta_{\alpha\beta} a_i, \tag{2.2c}$$

$$[b^\dagger_\alpha,\{a_i,a_j\}] = 0, \tag{2.2d}$$

$$[a^\dagger_i,\{b_\alpha,b_\beta\}] = 0. \tag{2.2e}$$

$$[b_\alpha,[b_\beta,b_\gamma]] = 0, \tag{2.3a}$$

$$[a_i,\{a_j,a_k\}] = 0, \tag{2.3b}$$

$$\{b_\alpha,\{b_\beta,a_i\}\} = [a_j,\{b_\beta,a_i\}] = 0, \tag{2.3c}$$

$$[b_\alpha,\{a_i,a_j\}] = [a_i,[b_\alpha,b_\beta]] = 0. \tag{2.3d}$$

Here b_α ($\alpha = 1,2,...,M$) and their conjugates $b^\dagger_\alpha$ are para-Fermi operators, and a_i ($i = 1,2,...,N$) and their conjugates $a^\dagger_i$ are para-Bose operators. The above relations are fully consistent with the Fermi–Bose interpretation of b_α and a_i. Note that though the relations (2.3) are independent of (2.1), the relations (2.2) can all be derived from the set (2.1) through generalized Jacobi identities.

The unmixed relations among b's [(2.1a), (2.2a), and (2.3a)] constitute the fundamental relations of para-Fermi statistics. Similarly the unmixed relations among a's [(2.1b), (2.2b), and (2.3b)] constitute the complete set of fundamental relations of para-Bose statistics. It is the remaining mixed ones among a's and b's that are new to the parastatistics we are considering in this work. Toward the end of this section we shall show that the fundamental relations of para-Fermi and para-Bose statistics arise from the

0022-2488/88/010016-05$02.50

"odd" and "even" sectors, whereas the remaining ones constitute their extension. First, however, we consider the following toy model of supersymmetric quantum mechanics in order to motivate the relations (2.1).

For simplicity let a and b be single para-Bose and para-Fermi operators, respectively. The fundamental relations among them and their conjugates are a special case of (2.1), (2.2), and (2.3), $M = N = 1$. Define the supergenerators as given by

$$Q = \tfrac{1}{2}\{a^\dagger,b\}, \tag{2.4a}$$

$$Q^\dagger = \tfrac{1}{2}\{b^\dagger,a\}, \tag{2.4b}$$

and the even generators as given by

$$H = \tfrac{1}{4}\{a^\dagger,a\} + \tfrac{1}{4}[b^\dagger,b]. \tag{2.5}$$

Using the above definitions of Q and $Q^\dagger$, which are constructed out of para-Bose and para-Fermi operators, it is easy to compute the brackets $\{Q,Q^\dagger\}$, $[Q,H]$, and $\{Q,Q\}$ and verify the algebraic structure (1.1) of supersymmetric quantum mechanics on using the identities given below and the fundamental relations (2.1a)–(2.1e). These identities, which hold for arbitrary operators Ω_1, Ω_2, Ω_3, and Ω_4, are

$$\{\{\Omega_1,\Omega_2\},\{\Omega_3,\Omega_4\}\} = \{\Omega_1,\{\Omega_2,\{\Omega_3,\Omega_4\}\}\} + [\Omega_2,[\Omega_1,\{\Omega_3,\Omega_4\}]], \tag{2.6a}$$

$$[\{\Omega_1,\Omega_2\},[\Omega_3,\Omega_4]] = \{\Omega_1,[\Omega_2,[\Omega_3,\Omega_4]]\} + \{\Omega_2,[\Omega_1,[\Omega_3,\Omega_4]]\}, \tag{2.6b}$$

$$[\{\Omega_1,\Omega_2\},\{\Omega_3,\Omega_4\}] = \{\Omega_1,[\Omega_2,\{\Omega_3,\Omega_4\}]\} + \{\Omega_2,[\Omega_1,\{\Omega_3,\Omega_4\}]\}, \tag{2.6c}$$

$$[[\Omega_1,\Omega_2,],[\Omega_3,\Omega_4]] = [\Omega_1,[\Omega_2,[\Omega_3,\Omega_4]]] - [\Omega_2,[\Omega_1,[\Omega_3,\Omega_4]]]. \tag{2.6d}$$

The full significance of the generalized supplementary relations (2.3), which are not required in the above derivation, emerges in the last section.

Before closing this section, let us dwell more on the generalization of the (unmixed) relations of ordinary parastatistics. Let (a_i, b_α) be denoted collectively by a_I, (a_j, b_β) by a_J, etc., where $I = (i,\alpha)$, $J = (j,\beta)$, etc., run over $M+N$ values. To distinguish a from b we introduce the odd/even "parity" $(-1)^{\Pi(I)}$ associated with a which equals -1 for the b-type operator and equals $+1$ for the a-type operator. Using this notation, (2.1)–(2.3) are compactly summarized to our advantage as

$$[a_I,\langle a_J^\dagger,a_K\rangle]_\pm = 2\delta_{IJ}a_K, \tag{2.7a}$$

$$[a_I^\dagger,\langle a_J,a_K\rangle]_\pm = -2(-1)^{\Pi(I)\Pi(J)}\delta_{IJ}a_K - 2(-1)^{\Pi(K)(\Pi(J)+\Pi(I))}\delta_{IK}a_J, \tag{2.7b}$$

$$[a_K,\langle a_L,a_M\rangle]_\pm = 0. \tag{2.7c}$$

Here $[\ ,\ ]_\pm$ denotes the generalized Lie bracket, e.g.,

$$[a_I,a_J]_\pm = a_Ia_J - (-1)^{\Pi(I)\Pi(J)}a_Ja_I,$$

and $\langle\ ,\ \rangle$ denotes the (anti-) symmetrized product, e.g.,

$$\langle a_I,a_J\rangle = a_Ia_J + (-1)^{\Pi(I)\Pi(J)}a_Ja_I.$$

These relations are all fully consistent with the canonical relations, which in our compact notation read

$$[a_I,a_J^\dagger]_\pm = -(-1)^{\Pi(I)\Pi(J)}[a_J^\dagger,a_I]_\pm = \delta_{IJ}, \tag{2.8a}$$

$$[a_I,a_J]_\pm = 0. \tag{2.8b}$$

The identities (2.6) can also be summed up into a single identity, which we would rather omit. The compact notation brings out clearly the "even" and "odd" sectors consisting of parabosons and parafermions. The fundamental relations that connect these two sectors are new to generalized parastatistics. In order to shed more light on them, we now turn to operator realizations of (graded) Lie algebras systematically.

III. OPERATOR REALIZATIONS OF THE UNITARY ALGEBRAS[3]

Operator realizations of the unitary algebras, using conventional Fermi/Bose-type of operators, are well known. The commutation relations are

$$[E_{ij},E_{kl}] = \delta_{jk}E_{il} - \delta_{il}E_{kj}. \tag{3.1}$$

Here $i,j,k,l = 1,2,\dots,N$. Let us introduce N Fermi/Bose operators b/a and their adjoints, which enjoy the fundamental relations

$$\{b_i^\dagger,b_j\} = \delta_{ij}, \quad \{b_i,b_j\} = 0, \tag{3.2a}$$

$$[a_i,a_j^\dagger] = \delta_{ij}, \quad [a_i,a_j] = 0, \tag{3.2b}$$

and the corresponding adjoint ones. Then the algebraic structure (3.1) is realized through the N^2 bilinear constructs given by

$$E_{ij} = b_i^\dagger b_j/a_i^\dagger a_j.$$

We observe that (3.1) is also tenable with the parafermionic/parabosonic interpretation of b_i/a_i and $b_i^\dagger/a_i^\dagger$. The kind of fundamental relations of para-Fermi/para-Bose statistics we require to verify this claim are (2.1a) and (2.1b), i.e.,

$$[b_j,[b_k^\dagger,b_l]] = 2\delta_{jk}b_l, \tag{3.3a}$$

$$[a_j,\{a_k^\dagger,a_l\}] = 2\delta_{jk}a_l, \tag{3.3b}$$

and the related adjoint ones. The commutation rules (3.1) follow directly if we identify E_{ij} with antisymmetrized/symmetrized bilinear constructs:

$$E_{ij} = \tfrac{1}{2}[b_i^\dagger,b_j]/\tfrac{1}{2}\{a_i^\dagger,a_j\}.$$

Hence the consistency of the structure (3.1) with parastatistics. In this derivation based on the fundamental relations of parastatistics it is convenient to use the identities (2.6c) and (2.6d). Note that we have only shown that (3.1) is consistent with the fundamental relations of parastatistics given by (3.3a)/(3.3b), which in turn are also consistent with (3.2a)/(3.2b) of ordinary statistics. But there are additional fundamental relations of parastatistics that, though they have not been used in arriving at the algebraic relations (3.1), are nonetheless consistent with (3.2a)/(3.2b) of ordinary statistics. Of these new relations, the ones that follow from (3.3a)/(3.3b) through generalized Jacobi identities are (2.2a) and (2.2b), i.e.,

$$[b_k^\dagger,[b_l,b_j]] = 2\delta_{kl}b_j - 2\delta_{kj}b_l, \qquad (3.4a)$$

$$[a_k^\dagger,\{a_l,a_j\}] = -2\delta_{kl}a_j - 2\delta_{kj}a_l, \qquad (3.4b)$$

together with the adjoint ones. The remaining ones are the famous supplementary fundamental relations (2.3a) and (2.3b), i.e.,

$$[b_k,[b_l,b_j]] = 0, \qquad (3.5a)$$

$$[a_k,\{a_l,a_j\}] = 0, \qquad (3.5b)$$

and their adjoint relations. In Secs. IV and V, we make use of the complete set of fundamental relations, [(3.3a), (3.4a), and (3.5a)]/[(3.3b), (3.4b), and (3.5b)], to show the consistency of operator realizations of orthogonal/symplectic algebras (which contain unitary algebras as their subalgebras) with parastatistics.

IV. PARA-FERMI STATISTICS AND OPERATOR REALIZATIONS OF ORTHOGONAL ALGEBRAS[4,5]

According to the Cartan classification of Lie algebras, the orthogonal algebras SO$(2N+1)$ and SO$(2N)$ are denoted by B_N and D_N, respectively. Given a Lie algebra of rank N, there are exactly N generators $H_1,H_2,\ldots,H_N$ that span its Cartan subalgebra. Since they all commute, they are chosen simultaneously diagonal in any linear representation of the algebra. The nth diagonal entry in H_i is the ith component of the weight associated with a given vector $|n\rangle$ in the linear vector space of any representation. The adjoint representation is of special importance because the $D-N$ distinct nonzero weights associated with it are the roots of the algebra whose order is D. For orthogonal algebras B_N and D_N of order $(2N+1)N$ and $N(2N-1)$, respectively, all the roots are conveniently enumerated in terms of N orthonormal vectors

$$e_1,e_2,\ldots,e_N$$

spanning N-dimensional Euclidean space. They are

$$B_N\colon\ \pm e_i \pm e_j, \pm e_i,$$

$$D_N\colon\ \pm e_i \pm e_j, \quad i>j=1,2,\ldots,N.$$

Each root α (components $\alpha_1,\alpha_2,\ldots,\alpha_N$) is associated with a ladder operator denoted by $E(\alpha)$. This is like a raising (lowering) operator, if α is a positive (negative) root. We now give the operator realizations for $E(\alpha)$, for every root α. With their help we derive the algebraic structure of a classical Lie algebra in its canonical form:

$$[E(\alpha),E(-\alpha)] = \sum_{i=1}^{N} \alpha_i H_i, \qquad (4.1a)$$

$$[H_i,E(\alpha)] = \alpha_i E(\alpha), \qquad (4.1b)$$

$$[E(\alpha),E(\beta)] = N_{\alpha\beta}E(\alpha+\beta). \qquad (4.1c)$$

Here $N_{\alpha\beta}$ is a c number that vanishes if $\alpha+\beta$ is not a root of the algebra. In the derivation of this canonical structure for orthogonal algebras, we might in the spirit of Sec. III start with the operators obeying Fermi statistics and then note the consistency of our derivation with para-Fermi statistics. However, in this section we prefer to work directly with para-Fermi operators

$$b_1,b_2,\ldots,b_N,$$

and their adjoints which enjoy the fundamental relations (3.3a), (3.4a), and (3.5a).

We associate the ladder operators $E(\pm e_i \pm e_j)$ corresponding to $\pm e_i \pm e_j$ with the following antisymmetrized bilinear constructs:

$$E(+e_i+e_j) = (i/2)[b_i^\dagger,b_j^\dagger],$$

$$E(-e_i-e_j) = (i/2)[b_i,b_j],$$

$$E(+e_i-e_j) = \tfrac{1}{2}[b_i^\dagger,b_j] = E(-e_j+e_i).$$

In the case of B_N, the ladder operators $E(\pm e_i)$ corresponding to $\pm e_i$ are associated with the elementary operators themselves:

$$E(+e_i) = b_i^\dagger/\sqrt{2}, \quad E(-e_i) = b_i/\sqrt{2}.$$

As a result of this prescription, let us examine what the commutation relations are. The identity (2.6d) proves very helpful. From (3.3a) we get the following commutation relations in accord with (4.1a) and (4.1b):

$$[E(+e_i),E(-e_i)] = H_i := [b_i^\dagger,b_i]/2,$$

$$[E(e_i-e_j),E(-e_i+e_j)] = H_i - H_j,$$

$$[H_i,E(\pm e_j)] = \pm\delta_{ij}E(\pm e_j),$$

$$[H_i,E(\pm e_j \pm e_k)] = (\pm\delta_{ij} \pm \delta_{ik})E(\pm e_j \pm e_k).$$

From (3.4a) we derive the following commutation relations in keeping with (4.1a):

$$[E(+e_i+e_j),E(-e_i-e_j)] = H_i + H_j.$$

Finally from (3.4a) and the supplementary relations (3.5a) we also get those commutation relations expected from (4.1c), e.g.,

$$[E(+e_i+e_j),E(e_k+e_l)] = 0,$$

$$[E(+e_i+e_j),E(-e_j+e_k)] = -E(e_i+e_k).$$

The above exercise reveals that the complete set of fundamental relations for parastatistics might also been discovered from the commutation relations for orthogonal algebras.

V. PARA-BOSE STATISTICS AND OPERATOR REALIZATIONS OF SYMPLECTIC ALGEBRA[4,5]

According to Cartan classifications of Lie algebras, the symplectic algebras Sp$(2N)$ are denoted by C_N. In terms of the set of N orthonormal vectors introduced in Sec. IV, all the roots of C_N are conveniently enumerated as

$$\pm e_i \pm e_j, \quad \pm 2e_i,$$

$i>j=1,2,\ldots N$. In this section we give the operator realizations for the ladder operators $E(\alpha)$, for every root α of C_N. This enables us to derive the algebraic structure of C_N in its canonical form:

$$[E(+2e_i),E(-2e_i)] = 2H_i, \qquad (5.1a)$$

$$[E(e_i-e_j),E(e_j-e_i)] = H_i - H_j, \qquad (5.1b)$$

$$[E(e_i+e_j),E(-e_j-e_i)] = H_i + H_j, \qquad (5.1c)$$

$$[H_i,E(\pm 2e_j)] = \pm 2\delta_{ij}E(\pm 2e_j), \qquad (5.1d)$$

$$[H_i,E(\pm e_j \pm e_k)] = (\pm\delta_{ij} \pm \delta_{ik})E(\pm e_j \pm e_k), \qquad (5.1e)$$

$$[E(\alpha),E(\beta)] = N_{\alpha\beta}E(\alpha+\beta). \tag{5.1f}$$

Here $N_{\alpha\beta}$ is a c number required to vanish when $\alpha+\beta$ is not a root of the algebra, e.g., if $\alpha = e_i + e_j$ and $\beta = e_k + e_l$. In the derivation of the structure (5.1a)–(5.1f) we might, at the outset, use the operators obeying Bose statistics and then in the spirit of Sec. III note the consistency of this structure with para-Bose statistics. However, in order to point out here that the complete set of fundamental relations for parastatistics might also have been discovered *en route* to the commutation rules for symplectic algebras, we prefer to work in this section directly with para-Bose operators $a_1,a_2,...,a_N$ and their adjoints. They enjoy the fundamental relations (3.3b), (3.4b), and (3.5b).

We give the following prescription:

$$\begin{aligned}
&E(e_i+e_j) = (i/2)\{a_i^\dagger,a_j^\dagger\}, \quad i>j,\\
&E(-e_i-e_j) = (i/2)\{a_i,a_j\}, \quad i>j,\\
&E(-e_i+e_j) = \tfrac{1}{2}\{a_i,a_j^\dagger\} = E(e_j-e_i), \quad i>j,\\
&E(+2e_i) = ia_i^{\dagger 2}/\sqrt{2}, \quad E(-2e_i) = ia_i^2/\sqrt{2}.
\end{aligned} \tag{5.2}$$

In order to justify the symplectic algebraic structure, the identity (2.6c) proves very helpful. From (3.3b), we derive (5.1b), (5.1d), and (5.1e). From (3.4b), we derive (5.1a) and (5.1e). The Cartan generators H are identified as

$$H_i = \tfrac{1}{2}\{a_i^\dagger,a_i\}.$$

Finally the verification of (5.1f) requires the use of supplementary relations (3.5b). In this way we find the consistency of the para-Bose statistics with the symplectic structure.

VI. GENERALIZED PARASTATISTICS AND OPERATOR REALIZATIONS OF GRADED-LIE ALGEBRAS

In this section, we generalize our considerations of Secs. IV and V to graded Lie algebras. The operator realizations of graded unitary algebras are already well known. They are conventionally based on M fermionic operators b_α and N bosonic operators a_i and their adjoints. We first observe that the resulting graded Lie structure gu (M/N) is entirely consistent with (2.1a)–(2.1e), i.e., with the generalized para-Fermi–Bose interpretation of b_α and a_i. Generalizing the considerations of the last two sections, we further note that the orthosymplectic structure is consistent with the complete set of postulated relations of generalized parastatistics given by (2.1)–(2.3), provided the following identifications are made for the even generators $E_{\alpha\beta}$, $E_{\bar\alpha\beta}$, $E_{\bar\alpha\bar\beta}$, E_{ij}, $E_{\bar\imath j}$, $E_{\bar\imath\bar\jmath}$, and the odd generators $S_{i\alpha}$, $S_{\bar\imath\alpha}$, $S_{\bar\alpha i}$, $S_{\bar\alpha\bar\imath}$ of the orthosymplectic algebra:

$$\begin{aligned}
&E_{\alpha\beta} = \tfrac{1}{2}[b_\alpha,b_\beta], \quad E_{\bar\alpha\beta} = \tfrac{1}{2}[b_\alpha^\dagger,b_\beta],\\
&E_{\bar\alpha\bar\beta} = \tfrac{1}{2}[b_\alpha^\dagger,b_\beta^\dagger],\\
&E_{ij} = \tfrac{1}{2}\{a_i,a_j\}, \quad E_{\bar\imath j} = \tfrac{1}{2}\{a_i^\dagger,a_j\}, \quad E_{\bar\imath\bar\jmath} = \tfrac{1}{2}\{a_i^\dagger,a_j^\dagger\},\\
&S_{i\alpha} = \tfrac{1}{2}\{a_i,b_\alpha\}, \quad S_{\bar\imath\alpha} = \tfrac{1}{2}\{a_i^\dagger,b_\alpha\},\\
&S_{\bar\alpha i} = \tfrac{1}{2}\{b_\alpha^\dagger,a_i\}, \quad S_{\bar\imath\bar\alpha} = \tfrac{1}{2}\{b_\alpha^\dagger,a_i^\dagger\}.
\end{aligned} \tag{6.1}$$

Given that the generators are thus constructed as bilinears of operators obeying generalized parastatistics, it is straightforward to work out the (anti-)commutation relations among them. We give some of them here. The bracket between an even and an odd generator is a commutator, e.g.,

$$\begin{aligned}
&[E_{\bar\alpha\bar\beta},S_{\bar\gamma i}] = 0, \quad [E_{\bar\alpha\bar\beta},S_{\gamma i}] = \delta_{\beta\gamma}S_{\bar\alpha i} - \delta_{\alpha\gamma}S_{\bar\beta i},\\
&[E_{\bar\imath\bar\jmath},S_{\bar\gamma k}] = \delta_{jk}S_{\bar\gamma\bar\imath} + \delta_{ik}S_{\bar\gamma\bar\jmath}, \quad [E_{\bar\imath\bar\jmath},S_{\bar\gamma\bar k}] = 0.
\end{aligned} \tag{6.2}$$

The bracket between two odd generators is an anticommutator, e.g.,

$$\{S_{\bar\alpha i},S_{\bar\beta j}\} = 0, \quad \{S_{\bar\alpha i},S_{\bar\jmath\beta}\} = \delta_{ij}S_{\bar\alpha\beta} + \delta_{\alpha\beta}E_{\bar\jmath i}. \tag{6.3}$$

Thus we find that generalized parastatistics is intimately related with orthosymplectic structure, as is ordinary parastatistics with orthogonal and symplectic structures.

The significant results in this work are the relative or mixed bracket relations, such as (2.1c)–(2.1e) and the algebraic constructions for the supergroup generators, such as (6.1). It is worth pointing out that the verification of the trilinear relations, such as (2.1)–(2.3), and also of the algebras of the generators of various groups, becomes simpler and almost self-evident if the operators a's and b's are decomposed into their Green's components and the group generators are also expressed in terms of them. Thus

$$a_i = \sum_A a_i^A, \quad b_\alpha = \sum_A b_\alpha^A. \tag{6.4}$$

Here the sum extends over all the Green's components, which are as many in number as the order of parastatistics obeyed by the operator. Following Greenberg and Messiah,[6] it can be shown that all the para-Bose and para-Fermi operators must have the same order to ensure validity of the nontrivial mixed trilinear relations.

In (6.4), we have normal relations

$$\begin{aligned}
[a_i^A,a_j^A]_- &= [a_i^A,a_j^{\dagger A}]_- - \delta_{ij}\\
&= [a_i^A,b_\alpha^A]_- = \{b_\alpha^A,b_\beta^A\}\\
&= \{b_\alpha^A,b_\beta^{\dagger A}\} - \delta_{\alpha\beta} = 0,
\end{aligned} \tag{6.5}$$

for each component given, and anomalous relations between two different components:

$$\begin{aligned}
\{a_i^A,a_j^B\} &= \{a_i^A,a_j^B\} = \{a_i^A,b_\alpha^B\}\\
&= [b_\alpha^A,b_\beta^B] = [b_\alpha^A,b_\beta^{B+}] = 0.
\end{aligned} \tag{6.6}$$

The algebraic constructions of the last three sections also become transparent through the introduction of Green's components. For example, in (6.1),

$$\begin{aligned}
&E_{\alpha\beta} = \sum_A \frac{1}{2}[b_\alpha^A,b_\beta^A], \quad E_{ij} = \sum_A \frac{1}{2}\{a_i^A,a_j^A\},\\
&S_{i\alpha} = \sum_A \frac{1}{2}\{a_i^A,b_\alpha^A\}.
\end{aligned} \tag{6.7}$$

Thus each generator constructed as a bilinear in para-Bose and para-Fermi operators reduces on their decomposition into Green's components into a direct sum of diagonal terms each of which is a bilinear in ordinary Bose–Fermi operators.

ACKNOWLEDGMENT

We are grateful to O. W. Greenberg for his helpful comments.

[1] E. Witten, Nucl. Phys. B **15**, 513 (1981).
[2] H. S. Green, Phys. Rev. **90**, 270 (1953).
[3] A. J. Bracken and H. S. Green, J. Math. Phys. **14**, 1784 (1973).
[4] C. Ryan and E. C. G. Sudarshan, Nucl. Phys. **47**, 207 (1963).
[5] P. Jordan, N. Mukunda, and S. V. Pepper, J. Math. Phys. **4**, 1089 (1963).
[6] O. W. Greenberg and A. M. L. Messiah, Phys. Rev. B **138**, 1155 (1965).

The analogy between spin glasses and Yang–Mills fluids

Darryl D. Holm
Center for Nonlinear Studies and Theoretical Division, MS B284, Los Alamos National Laboratory, Los Alamos, New Mexico 87545

Boris A. Kupershmidt
University of Tennessee Space Institute, Tullahoma, Tennessee 87388

(Received 21 July 1987; accepted for publication 19 August 1987)

A dictionary of correspondence is established between the dynamical variables for spin-glass fluid and Yang-Mills plasma. The Lie-algebraic interpretation of these variables is presented for the two theories. The noncanonical Poisson bracket for the Hamiltonian dynamics of an ideal spin glass is shown to be identical to that for the dynamics of a Yang–Mills fluid plasma, although the Hamiltonians differ for the two theories. This Poisson bracket is associated to the dual space of an infinite-dimensional Lie algebra of semidirect-product type.

I. INTRODUCTION

A. Background physics

Halperin and Saslow[1] and Andreev[2] have introduced condensed-matter theories of spin glasses, i.e., disordered magnetic spin systems whose ground states are degenerate under rotations. Condensed-matter systems whose ground states are degenerate under a continuous symmetry are often described macroscopically in terms of an order parameter field taking values in the Lie group associated with that symmetry. The order parameter field describing the low-temperature configurations of a spin glass in the Halperin–Saslow–Andreev theory is a spatially varying orthogonal matrix $O(\mathbf{x})$, acting on classical spin vectors at each point $\mathbf{x}$. The matrix $O(\mathbf{x})$ is assumed to be slowly varying in space (see also Toulouse,[3] Henley *et al.*,[4] Bray and Moore,[5] Saslow,[6] and Henley[7]). The spins themselves may be eliminated in the Halperin–Saslow–Andreev theory and their dynamics replaced by that of the order parameter field, $O(\mathbf{x},t)$.

Singularities in the order parameter field are called defects. These defects can be classified topologically by conventional homotopy theory (Toulouse and Kléman,[8] Volovik and Mineev,[9] Mermin,[10] and Michel[11]). The presence of defects (singularities in the order parameter field) suggests introducing additional degrees of freedom that may be described by gauge fields associated to the symmetry group of the degenerate ground state. For spin glasses, the symmetry group is SO(3) and these additional gauge fields have been introduced heuristically (in Dzyaloshinskii and Volovik,[12,13] Hertz,[14] José and Hertz,[15] and Dzyaloshinskii[16,17]) by replacing ordinary space derivatives with covariant derivatives according to the SO(3) minimal-coupling prescription in Hamilton's principle at the level of Ginzburg–Landau mean field theory for the order-parameter dynamics. (See also Fischer[18] and Rozhkov[19].) This Ginzburg–Landau type of model could presumably be derived from a lattice model in three dimensions (by the renormalization group method, for example), but as yet no explicit connection seems to have been made between the macroscopic gauge fields and microscopic concepts such as frustration in more than two dimensions. For the two-dimensional case, the concept of local exchange invariance on a frustrated planar lattice leads naturally to an analogy between nonlinear spin-glass hydrodynamics and Yang–Mills SO(3) gauge theory (see, e.g., Refs. 12 and 13).

The Ginzburg–Landau theory with covariant derivatives describes the dynamics of isolated defects in terms of dynamics of a gauge field. Interactions among defects (Andreev[2]) and defect cores (Kawasaki and Brand[20]) may be introduced by modifying the Hamiltonian or free energy of the system. For the case of spin glasses, the phenomenological theory so defined lacks the couplings between space and spin indices that complicate the free energies of superfluid ^{3}He-B (Toulouse and Kléman[8]), and cholesteric liquid crystals (Toulouse and Kléman,[8] Bouligand *et al.*,[21] Mermin[10]), which can also be described by order parameter fields taking values in SO(3). Other generalizations also exist, such as (1) local anisotropy (Saslow[22]), (2) remanence, an external field, or a tendency toward ferromagnetism (Halperin and Saslow[1]), and (3) dissipation, e.g., spin diffusion and relaxation of the order parameter (Halperin and Saslow[1]). Recent reviews of spin glasses are given in Fischer[18] and Chowdhury and Mookerjee.[23]

B. Problem statement

As one can glean from the previous remarks, there exists at least a partial analogy between fluid dynamics with internal degrees of freedom (e.g., spin-glass dynamics, superfluids, and other quantum liquids) and Yang–Mills fluid dynamics. This analogy was introduced for spin glasses by Dzyaloshinskii and Volovik[12,13] and Volovik and Dotsenko,[24] and discussed for superfluids and other quantum liquids by Dzyaloshinskii and Volovik[13] and Khalatnikov and Lebedev.[25] Here, we propose to examine this analogy in the framework of the Hamiltonian formulation of nonlinear hydrodynamic theories. In Sec. II, we present a unification of the nondissipative theories of spin-glass dynamics, Yang–Mills plasmas, and Yang–Mills magnetohydrodynamics that combines their various Hamiltonian formulations into a single Poisson bracket, which we associate in Sec. III to the dual space of a Lie algebra endowed with two different types of nontrivial generalized two cocycles.

During the past few years, Poisson bracket methods have been used to derive nonlinear hydrodynamic equations for various complex fluid systems. These systems include spin glasses (Dzyaloshinskii and Volovik[13]); hydrodynamics of defects in the continuum description of condensed matter, e.g., vortices in superfluid ^{4}He and disclinations in a planar magnet (Volovik and Dotsenko[24]); rotating superfluid ^{4}He and ^{3}He with spin and orbital angular momentum (Khalatnikov and Lebedev,[25] Holm and Kupershmidt[26]); as well as Yang–Mills plasmas (Gibbons, Holm, and Kupershmidt,[27] Holm and Kupershmidt[28]).

The Poisson bracket method provides a guide for determining conservation laws and a framework for studying Lyapunov stability of equilibrium solutions (see Holm *et al.*[29]), as well as a structure for pointing out similarities and differences among various theories. We emphasize the latter structural aspect in this work, by showing that the Poisson brackets for spin glasses and Yang–Mills plasmas are isomorphic. Thus, although the Hamiltonians and physical interpretations of the two theories differ, the Lie-algebraic nature of their Hamiltonian structures is the same. This Lie-algebraic nature allows us in Sec. IV to set up a dictionary of correspondence between the dynamical variables for spin-glass fluid and Yang–Mills plasma.

II. SPIN-GLASS DYNAMICS AND YANG–MILLS MAGNETOHYDRODYNAMICS

The gauge-field formulation of the nonlinear hydrodynamic equations describing the continuum dynamics of defects in condensed matter is developed in Dzyaloshinskii and Volovik.[12,13,30] In this formulation, gauge fields are introduced via the minimal-coupling hypothesis in Hamilton's principle as additional variables coupled to the defects, represented in turn as densities of gauge charges. Physical applications include crystals with continuously distributed dislocations and disclinations; superfluid HeII with vortices; liquid crystals with rotational disclinations; and two-dimensional spin glasses, regarded as the continuum limit of a planar lattice of magnets with disclinations.

The problem of formulating nonlinear dynamical equations for ideal (nondissipative) media containing continuously distributed defects is addressed here via the Hamiltonian approach. That is, the dynamics of a continuously defected medium is represented in Hamiltonian form, i.e., as

$$\partial_t \mathbf{u} = \{H,\mathbf{u}\}, \tag{2.1}$$

for Hamiltonian H and Poisson bracket $\{\ ,\ \}$ defined on the space of dynamical variables $\mathbf{u}$.

An example of such a system and the starting point for the present analysis is the theory for spin glass (continuum limit of an antiferromagnet having nonzero equilibrium disclination density) of Volovik and Dotsenko.[24] In this theory, the gauge-charge density G is the three-component magnetization density, which generates the internal symmetry group of three-dimensional rotations. The corresponding gauge potential A_i, $i = 1,2,3$, transforms under these internal symmetry rotations like a gauge field (see, e.g., Drechsler and Mayer[31]). The disclination density is identified with the gauge-field intensity

$$B_{ik} = A_{i,k} - A_{k,i} + [A_i, A_k] \tag{2.2a}$$

or, componentwise,

$$B^{\alpha}_{ik} = A^{\alpha}_{i,k} - A^{\alpha}_{k,i} + t^{\alpha}_{\beta\gamma} A^{\beta}_{i} A^{\gamma}_{k}, \tag{2.2b}$$

with notation explained below.

In our notation, Latin indices $i,j,k,\ldots$, run from 1 to n ($n=3$ for three-dimensional space), script Latin indices $a,b,c,\ldots$, run from 0 to n, and the charge G belongs to the dual $\mathfrak{g}^*$ of the gauge-symmetry Lie algebra $\mathfrak{g}$, with $A_i \in \mathfrak{g}$. The adjoint representation map ad: $\mathfrak{g} \to \operatorname{End} \mathfrak{g}$ denotes multiplication in $\mathfrak{g}$: $\operatorname{ad}(y)z = [y,z]$. Another map $\mathfrak{g} \to \mathfrak{g}^*$, $\mathfrak{g} \ni y \to {}^*y \in \mathfrak{g}^*$ is defined by the rule

$$\langle {}^*y, z \rangle = (y,z), \tag{2.3}$$

where $(\ ,\ )$ is an invariant symmetric nondegenerate form on $\mathfrak{g}$ (e.g., the Killing form, for $\mathfrak{g}$ semisimple). The structure constants of $\mathfrak{g}$ are denoted $t^{\gamma}_{\alpha\beta}$ [see (2.2)] in a basis with elements e_α, where Greek indices run from 1 to $M = \dim \mathfrak{g}$. In this basis, we have the commutator relation

$$[e_\alpha, e_\beta] = t^{\gamma}_{\alpha\beta} e_\gamma. \tag{2.4}$$

We denote $A_i = A^{\alpha}_{i} e_\alpha$ and $G = G_\alpha e^\alpha$, where e^α, $\alpha = 1,\ldots,M$, are elements of the dual basis, satisfying $\langle e^\alpha, e_\beta \rangle = \delta^\alpha_\beta$. The rule (2.3) associates to each element $y \in \mathfrak{g}$ a corresponding dual element ${}^*y \in \mathfrak{g}^*$, via

$${}^*y_\beta z^\beta := \langle {}^*y, z \rangle = (y,z) =: y^\alpha g_{\alpha\beta} z^\beta, \tag{2.5}$$

where $g_{\alpha\beta} = (e_\alpha, e_\beta)$ is the matrix of the invariant form in the basis $\{e_\alpha\}$.

To the linear operation ad on $\mathfrak{g}$, there corresponds another linear operation ad* (essentially minus the transpose of ad, in a matrix representation), which acts on $\mathfrak{g}^*$ as defined by

$$\langle \operatorname{ad}^*(y)\,{}^*z, x \rangle := \langle {}^*z, [x,y] \rangle \tag{2.6}$$

for $x,y \in \mathfrak{g}$ and ${}^*z \in \mathfrak{g}^*$. In components, then,

$$\begin{aligned}(\operatorname{ad}^*(y)\,{}^*z)_\alpha x^\alpha &= \langle \operatorname{ad}^*(y)\,{}^*z, x \rangle = -\langle {}^*z, \operatorname{ad}(y)x \rangle \\ &= -\langle {}^*z, [y,x] \rangle = -z_\gamma t^{\gamma}_{\beta\alpha} y^\beta x^\alpha,\end{aligned}$$

so that

$$(\operatorname{ad}^*(y)\,{}^*z)_\alpha = -y^\beta t^{\gamma}_{\beta\alpha} z_\gamma. \tag{2.7}$$

We may now define $(n+1)$ covariant derivative operators acting on $\mathfrak{g}$-valued functions of space and time. Namely,

$$\mathbf{D} = \nabla - \operatorname{ad}(\mathbf{A}), \tag{2.8a}$$

with n spatial components

$$D_i = \partial_i - \operatorname{ad}(A_i), \tag{2.8b}$$

and

$$D_t = \partial_t - \operatorname{ad}(A_0), \tag{2.9}$$

for the time component. Similarly, one defines $(n+1)$ covariant derivative operators acting on $\mathfrak{g}^*$-valued functions

$$\mathbf{D}^* = \nabla - \operatorname{ad}^*(\mathbf{A}), \tag{2.10a}$$

with components

$$D^*_i = \partial_i - \operatorname{ad}^*(A_i), \tag{2.10b}$$

and

$$D^*_t = \partial_t - \operatorname{ad}^*(A_0). \tag{2.11}$$

If ϕ and $\bar{\phi}$ are functions of space and time with values in $\mathfrak{g}$ and

$\mathfrak{g}^*$, respectively, then, for example, we have the partial-derivative relations

$$\langle\bar{\phi},\phi\rangle_{,i} = \langle(D_i^*\bar{\phi}),\phi\rangle + \langle\bar{\phi},(D_i\phi)\rangle, \tag{2.12}$$

since

$$D_i^* = -(D_i)^\dagger, \tag{2.13}$$

where † stands for the "adjoint."

We also have

$$*[D_a(w)] = D_a^*(*w), \quad \forall w\in\mathfrak{g}, \quad \forall a\in(0,\dots,n). \tag{2.14}$$

Indeed, for any $y\in\mathfrak{g}$, we have, denoting $x = A_a$,

$$\begin{aligned}\langle *[D_a(w)],y\rangle &= (D_a(w),y) = (w_{,a} - \mathrm{ad}_x(w),y)\\ &= (w_{,a},y) - ([x,w],y)\\ &= (w_{,a},y) + (w,[x,y]),\end{aligned} \tag{2.15a}$$

and

$$\begin{aligned}\langle D_a^*(*w),y\rangle &= \langle *w_{,a},y\rangle - \langle \mathrm{ad}_x^*(*w),y\rangle\\ &= (w_{,a},y) + \langle *w,[x,y]\rangle\\ &= (w_{,a},y) + (w,[x,y]).\end{aligned} \tag{2.15b}$$

Comparison of (2.15a) and (2.15b) proves (2.14).

From the covariant derivative operators, one defines the fields

$$[D_i,D_t] = \mathrm{ad}(E_i), \tag{2.16a}$$

$$[D_i,D_j] = \mathrm{ad}(B_{ij}), \tag{2.16b}$$

with spatial components

$$E_i = A_{i,t} - A_{0,i} + [A_i,A_0] = F_{0i} = -F_{i0}, \tag{2.17a}$$

$$B_{ij} = A_{i,j} - A_{j,i} + [A_i,A_j] = -F_{ij}, \tag{2.17b}$$

where subscript-comma notation is used for partial derivatives, e.g., $A_{0,i} = (\partial A_0/\partial x^i)$. In n spatial dimensions, the one-form E has n spatial components E_i, and the two-form B has $n(n-1)/2$ independent spatial components B_{ij}, with $B_{ij} = -B_{ji}$ (skew symmetric).

In Yang–Mills plasma theory (Gibbons, Holm, and Kupershmidt,[27] Holm and Kupershmidt[28]), the Yang–Mills fields F_{ab} appearing in (2.17) satisfy

$$*(D_a F^{ab}) = J^b, \quad a,b, = 0,1,\dots,n, \tag{2.18}$$

where J^b with components $J^0 = G$, $J^i = Gv^i$, is the gauge current density, with v^i, $i = 1,\dots,n$, denoting velocity components of the moving medium. Script indices are raised and lowered by the Lorentz metric, with signature $(n-1)$. The gauge charge is conserved, since

$$\begin{aligned}D_b^* J^b &= D_b^{*}{*}(D_a F^{ab}) \quad [\text{by } (2.18)]\\ &= *(D_b D_a F^{ab}) \quad [\text{by } (2.14)]\\ &= -\tfrac{1}{2}\mathrm{ad}(F_{ab})F^{ab} = 0 \quad [\text{by (2.16) and (2.17)}].\end{aligned} \tag{2.19}$$

In the Volovik–Dotsenko spin-glass theory, the structure constants $t^\alpha_{\beta\gamma}$ in (2.2) for the gauge symmetry algebra are those of so(3): $t^\alpha_{\beta\gamma} = \epsilon_{\alpha\beta\gamma}$, the totally antisymmetric tensor in dim $\mathfrak{g} = 3$ dimensions, with $\epsilon_{123} = -1$. Let K_i be the defect momentum density and ρ the inertial mass density of the defects. The Poisson bracket for spin glass proposed by Volovik and Dotsenko[24] is then expressible as

$$\begin{aligned}\{H,F\} = -\int d^n x\bigg\{&\frac{\delta F}{\delta K_j}\left[(K_i\partial_j + \partial_i K_j + B^\alpha_{ji}G_\alpha)\frac{\delta H}{\delta K_i} + B^\alpha_{ji}\frac{\delta H}{\delta A^\alpha_i} + \rho\partial_j\frac{\delta H}{\delta \rho}\right]\\ &+ \frac{\delta F}{\delta G_\alpha}\left[t^\gamma_{\alpha\beta}G_\gamma\frac{\delta H}{\delta G_\beta} + (\partial_i\delta^\beta_\alpha - t^\beta_{\alpha\gamma}A^\gamma_i)\frac{\delta H}{\delta A^\beta_i}\right] + \frac{\delta F}{\delta\rho}\partial_i\rho\frac{\delta H}{\delta K_i}\\ &+ \frac{\delta F}{\delta A^\alpha_j}\left[B^\alpha_{ji}\frac{\delta H}{\delta K_i} + (\delta^\alpha_\beta\partial_j + t^\alpha_{\beta\gamma}A^\gamma_j)\frac{\delta H}{\delta G_\beta}\right]\bigg\},\end{aligned} \tag{2.20}$$

in three dimensions $(n = 3)$ and for functionals H and F of the dynamical variables $(K_j, G_\alpha, \rho, A^\alpha_j)$. In Hamiltonian matrix form, the spin-glass equations corresponding to the Poisson bracket (2.20) are

$$\partial_t\begin{vmatrix}K_j\\ \rho\\ A^\alpha_j\\ G_\alpha\end{vmatrix} = -\begin{vmatrix}(K_i\partial_j + \partial_i K_j + B^\alpha_{ji}G_\alpha) & \rho\partial_j & B^\beta_{ji} & 0\\ \partial_i\rho & 0 & 0 & 0\\ B^\alpha_{ji} & 0 & 0 & (\delta^\alpha_\beta\partial_j + t^\alpha_{\beta\gamma}A^\gamma_j)\\ 0 & 0 & (\delta^\beta_\alpha\partial_i - t^\beta_{\alpha\gamma}A^\gamma_i) & t^\gamma_{\alpha\beta}G_\gamma\end{vmatrix}\begin{vmatrix}\delta H/\delta K_i\\ \delta H/\delta\rho\\ \delta H/\delta A^\beta_i\\ \delta H/\delta G_\beta\end{vmatrix} \tag{2.21a}$$

for Hamiltonian [Volovik and Dotsenko,[24] Eq. (6.11)]

$$H = \int d^n x\left[\frac{1}{2\rho}|\mathbf{K}|^2 + \frac{1}{2}\rho{*}\mathbf{A}_\alpha\cdot\mathbf{A}^\alpha + \frac{1}{2\chi}G^*_\alpha G^\alpha\right], \tag{2.21b}$$

with constant susceptibility χ.

In general, Hamiltonian equations are expressible as

$$\partial_t\mathbf{u} = \mathsf{b}\cdot\frac{\delta H}{\delta\mathbf{u}} = \{H,\mathbf{u}\}, \tag{2.22}$$

where the Hamiltonian matrix b defines the Poisson bracket $\{H,F\}$ in terms of the dynamical variables $\mathbf{u}$ according to the standard form

$$\{H,F\} = \int d^nx \frac{\delta F}{\delta \mathbf{u}} \cdot \mathsf{b} \cdot \frac{\delta H}{\delta \mathbf{u}}. \tag{2.23}$$

The spin-glass Poisson bracket in (2.20) defined by the Hamiltonian matrix b given in (2.21a) is bilinear, skew symmetric, and satisfies the Jacobi identity. To demonstrate the last property (which is neither self-evident nor trivial), we map the Hamiltonian matrix b in (2.21a) into *an affine* form, by using the invertible transformation

$$\mathbf{P} = \mathbf{K} + G_\alpha \mathbf{A}^\alpha, \tag{2.24}$$

and leaving ρ, $\mathbf{A}^\alpha$, and G_α unchanged. Under such a map, the Hamiltonian matrix b changes according to the chain rule, i.e.,

$$\mathsf{b}_1 = \mathsf{J}\cdot\mathsf{b}\cdot\mathsf{J}^\dagger, \tag{2.25}$$

where J is the Fréchet derivative of the map (2.24) and $\mathsf{J}^\dagger$ is its adjoint. The resulting Hamiltonian formulation of the spin-glass equations in the new variables $(P_i,\rho,A_i^\alpha,G_\alpha)$ is found [after matrix multiplication as in (2.25) and elimination of old variables **K** in favor of new ones **P**], to be

$$\partial_t \begin{vmatrix} P_i \\ \rho \\ A_i^\alpha \\ G_\alpha \end{vmatrix} = - \begin{vmatrix} P_k\partial_i + \partial_k P_i & \rho\partial_i & \partial_k A_i^\beta - A_{k,i}^\beta & G_\beta \partial_i \\ \partial_k \rho & 0 & 0 & 0 \\ A_k^\alpha \partial_i + A_{i,k}^\alpha & 0 & 0 & \delta_\beta^\alpha \partial_i + t_{\beta\gamma}^\alpha A_i^\gamma \\ \partial_k G_\alpha & 0 & \delta_\alpha^\beta \partial_k - t_{\alpha\gamma}^\beta A_k^\gamma & t_{\alpha\beta}^\gamma G_\gamma \end{vmatrix} \begin{vmatrix} \delta H/\delta P_k \\ \delta H/\delta \rho \\ \delta H/\delta A_k^\beta \\ \delta H/\delta G_\beta \end{vmatrix}, \tag{2.26a}$$

with Hamiltonian

$$H = \int d^nx \left[\frac{1}{2\rho} |\mathbf{P} + G_\alpha \mathbf{A}^\alpha|^2 + \frac{1}{2}\rho^* \mathbf{A}_\alpha \cdot \mathbf{A}^\alpha + \frac{1}{2\chi} G_\alpha^* G^\alpha \right]. \tag{2.26b}$$

By being affine (linear plus constant) in the dynamical variables, the Hamiltonian matrix b_1 in (2.26a) yields a Poisson bracket [given by (2.23) with b replaced by b_1] that may be associated to the dual space of a certain Lie algebra with a generalized two-cocycle on it (Kupershmidt[32]). In this case, the Lie algebra is of semidirect-product type,

$$\mathfrak{g}_1 = D \circledS \{\Lambda^0 \oplus [(\Lambda^0 \otimes \mathfrak{g}) \circledS (\Lambda^{n-1} \otimes \mathfrak{g}^*)]\}, \tag{2.27}$$

where D is the Lie algebra of vector fields on $\mathbf{R}^n$ and Λ^i is the space of differential i forms on $\mathbf{R}^n$. The dual coordinates are P_i dual to $\partial_i \in D$; ρ, to $1 \in \Lambda^0$; G_α, to $1 \otimes e^\alpha \in (\Lambda^0 \otimes \mathfrak{g})$, i.e., functions taking values in Lie algebra $\mathfrak{g}$, the symmetry algebra; and A_i^α dual to $(\partial_i \lrcorner\, d^nx) \otimes e^\alpha \in (\Lambda^{n-1} \otimes \mathfrak{g}^*)$, i.e., $(n-1)$ forms taking values in the dual symmetry algebra $\mathfrak{g}^*$. In (2.27), $\circledS$ denotes semidirect product; $\otimes$, tensor product; and $\oplus$, direct sum. Mathematical discussion of this Lie algebra is deferred until Sec. III. At this point, we only remark that association of the b_1 Poisson bracket to the dual of the Lie algebra $\mathfrak{g}_1$ assures that the Jacobi identity for the b_1 Poisson bracket is satisfied. Since b_1 is related to b in (2.21a) by the invertible transformation (2.24), the Jacobi identity is also satisfied for the Poisson bracket (2.20) defined by the Hamiltonian matrix b in (2.21a).

Remarkably enough, a gauge-covariant Poisson bracket for spin glasses exists and is canonically related via (2.17b) to the Poisson bracket corresponding to the Hamiltonian matrix b_1 expressed in (2.26a) in terms of gauge potential A_i^α. The new Hamiltonian matrix b_2 is expressed in terms of the gauge field (disclination density) B_{ij}^α, using definition (2.2) in a chain-rule matrix multiplication as in (2.25). The resulting matrix-Hamiltonian equations for spin glass are now, in terms of B_{ij}^α [cf. (2.26a)],

$$\partial_t \begin{vmatrix} P_i \\ \rho \\ B_{ij}^\alpha \\ G_\alpha \end{vmatrix} = - \begin{vmatrix} P_k\partial_i + \partial_k P_i & \rho\partial_i & -B_{lm,i}^\beta + \partial_m B_{li}^\beta - \partial_l B_{mi}^\beta & G_\beta \partial_i \\ \partial_k \rho & 0 & 0 & 0 \\ B_{ij,k}^\alpha + B_{ik}^\alpha \partial_j - B_{jk}^\alpha \partial_i & 0 & 0 & t_{\beta\gamma}^\alpha B_{ij}^\gamma \\ \partial_k G_\alpha & 0 & -t_{\alpha\gamma}^\beta B_{lm}^\gamma & t_{\alpha\beta}^\gamma G_\gamma \end{vmatrix} \begin{vmatrix} \delta H/\delta P_k \\ \delta H/\delta \rho \\ \delta H/\delta B_{lm}^\beta \\ \delta H/\delta G_\beta \end{vmatrix}. \tag{2.28}$$

The Hamiltonian matrix b_2 in (2.28) is now *linear* in the dynamical variables and thus (Kupershmidt[32]) may be associated to the dual of a Lie algebra. In this case, the Lie alebra is again a semidirect product,

$$\mathfrak{g}_3 = D \circledS \{\Lambda^0 \oplus [(\Lambda^0 \otimes \mathfrak{g}) \circledS (\Lambda^{n-2} \otimes \mathfrak{g}^*)]\}, \tag{2.29}$$

with the *same* dual coordinates as in the case of b_1 associated to $\mathfrak{g}_1$in (2.27) *except* that instead of A_i^α dual to $(\partial_i \lrcorner\, d^nx) \otimes e^\alpha \in (\Lambda^{n-1} \otimes \mathfrak{g}^*)$, we now have B_{ij}^α dual to $(\partial_i \lrcorner\, \partial_j \lrcorner\, d^nx) \otimes e^\alpha \in (\Lambda^{n-2} \otimes \mathfrak{g}^*)$, i.e., $\{B_{ij}^\alpha\}$ dual to $(n-2)$ forms taking values in the dual gauge algebra, $\mathfrak{g}^*$.

Yang–Mills MHD: The Poisson matrices b_1 and b_2 for spin glasses in (2.26a) and (2.28) extend the corresponding matrices for Yang–Mills magnetohydrodynamics (YM–MHD) (Holm and Kupershmidt[28]), by allowing nonzero entries for Poisson brackets between the gauge charges and the gauge fields. The Hamiltonian for YM–MHD is (Holm and Kupershmidt[28])

$$H=\int d^{n}x\left[\frac{1}{2\rho}\,|\mathbf{P}|^{2}+U(\rho)+\frac{1}{4}({}^{*}B_{\alpha ij}B^{\alpha}_{ij})\right]. \tag{2.30}$$

Remarkably, when the YM–MHD Hamiltonian (2.30) is used with the *spin-glass* Hamiltonian matrices $\mathfrak{b}_1$ and $\mathfrak{b}_2$ in (2.26a) and (2.28), respectively, the same dynamical equations reemerge for YM–MHD as in Holm and Kupershmidt.[28] That is, correct YM–MHD equations reappear using the spin-glass Poisson bracket (2.28) with the YM–MHD Hamiltonian (2.30).

The spin-glass Hamiltonian matrices $\mathfrak{b}_1$ and $\mathfrak{b}_2$ extend their YM–MHD counterparts found in Holm and Kupershmidt[28] by allowing semidirect-product actions instead of simple direct sums among quantities dual to gauge charges and gauge fields, in the Lie algebras $\mathfrak{g}_1$ and $\mathfrak{g}_3$. Since this extension of Hamiltonian matrices is available for YM–MHD, it is natural to expect the Hamiltonian matrix for chromohydrodynamics (CHD: the non-Abelian Yang–Mills plasma theory from which YM–MHD is derived) also to have an extended counterpart. This extended counterpart may, in turn, find application in the theory of condensed matter with internal symmetry variables.

To determine this extension of the CHD Hamiltonian matrix, we propose to argue heuristically: we start from the extended YM–MHD/spin-glass Hamiltonian matrix in (2.26a) and enlarge it, by comparing its structure to that for Abelian charged fluids (Holm[33]).

There is a standard derivation (see, e.g., Friedberg[34]) of *Abelian* MHD from the ideal two-fluid Abelian plasma equations. Abelian MHD emerges in the course of this derivation in the limit that the dielectric constant vanishes (i.e., displacement current is neglected), the inertia of one species (the electrons) is negligible compared to the other (the ions), local charge neutrality is imposed, and drift effects (diamagnetic and Hall electric fields) are neglected. In Holm and Kupershmidt[28] this derivation has been adapted for the purpose of obtaining the non-Abelian YM–MHD theory from the equations of chromohydrodynamics (CHD), treated in Gibbons, Holm, and Kupershmidt.[27] The CHD equations describe non-Abelian Yang–Mills plasma theory, e.g., quark-gluon plasma physics, in the fluid description obtained by taking moments of the corresponding kinetic theory with particles interacting via Yang–Mills fields (i.e., Wong's equations). A consistent Hamiltonian theory of special relativistic CHD also exists (Holm and Kupershmidt[28]).

Abelian MHD may also be considered as a special case of the Hamiltonian theory of Abelian charged-fluid (ACF) motion that includes moving-material electromagnetic effects. The equations of ACF dynamics are given in the following Hamiltonian matrix form in Holm[33]:

$$\partial_t\left|\begin{array}{c} P_i \\ \rho \\ A_i \\ {}^{*}E_i \\ Q \end{array}\right| = -\left|\begin{array}{ccccc} P_k\partial_i+\partial_k P_i & \rho\partial_i & \partial_k A_i - A_{k,i} & {}^{*}E^{k}\partial_i - \partial_j^{*}E^{j}\delta_i^k & Q\partial_i \\ \partial_k\rho & 0 & 0 & 0 & 0 \\ A_k\partial_i + A_{i,k} & 0 & 0 & \delta_i^k & s\partial_i \\ \partial_k{}^{*}E^{i} - {}^{*}E^{j}\partial_j\delta_k^i & 0 & -\delta_k^i & 0 & 0 \\ \partial_k Q & 0 & s\partial_k & 0 & 0 \end{array}\right| \left|\begin{array}{c} \delta H/\delta P_k \\ \delta H/\delta\rho \\ \delta H/\delta A_k \\ \delta H/\delta{}^{*}E^{k} \\ \delta H/\delta Q \end{array}\right| . \tag{2.31}$$

In (2.31) the Abelian charge density Q satisfies Gauss's law,

$$a\rho = Q = \operatorname{div}{}^{*}\mathbf{E}, \tag{2.32}$$

which is preserved by the dynamics. In (2.32) the quantity a is the constant charge-to-mass ratio in ACF dynamics and $^{*}\mathbf{E}$ is the electric displacement vector. Actually, the Hamiltonian matrix in (2.31) is a slight extension of that in Holm[33] to include a generalized two-cocycle between Q and $\mathbf{A}$ [the terms proportional to the arbitrary constant s in (2.31)]; Holm[33] chooses $s=0$.

The Hamiltonian matrix for Abelian MHD may be recovered either from (2.31) for $s=1$ when the displacement vector $^{*}\mathbf{E}$ is absent, or from (2.26a) in the Abelian case, when the structure constants $t^{\alpha}_{\beta\gamma}$ vanish.

Comparing the Hamiltonian matrices (2.26a) for non-Abelian MHD and (2.31) for Abelian charged fluids suggests the following Hamiltonian matrix for the dynamics of *non-Abelian* charged fluids:

$$\partial_t\left|\begin{array}{c} P_i \\ \rho \\ A^{\alpha}_{i} \\ {}^{*}E^{i}_{\alpha} \\ G_\alpha \end{array}\right| = -\left|\begin{array}{ccccc} P_k\partial_i+\partial_k P_i & \rho\partial_i & \partial_k A^{\beta}_{i} - A^{\beta}_{k,i} & {}^{*}E^{k}_{\beta}\partial_i - \partial_j^{*}E^{j}_{\beta}\delta_i^k & G_\beta\partial_i \\ \partial_k\rho & 0 & 0 & 0 & 0 \\ A^{\alpha}_{k}\partial_i + A^{\alpha}_{i,k} & 0 & 0 & \delta^{\alpha}_{\beta}\delta_i^k & s\delta^{\alpha}_{\beta}\partial_i + t^{\alpha}_{\beta\gamma}A^{\gamma}_{i} \\ \partial_k{}^{*}E^{i}_{\alpha} - {}^{*}E^{j}_{\alpha}\partial_j\delta_k^i & 0 & -\delta^{\beta}_{\alpha}\delta_k^i & 0 & t^{\gamma}_{\alpha\beta}\,{}^{*}E^{i}_{\gamma} \\ \partial_k G_\alpha & 0 & s\delta^{\beta}_{\alpha}\delta_k - t^{\beta}_{\alpha\gamma}A^{\gamma}_{k} & t^{\gamma}_{\alpha\beta}\,{}^{*}E^{k}_{\gamma} & t^{\gamma}_{\alpha\beta}G_\gamma \end{array}\right| \left|\begin{array}{c} \delta H/\delta P_k \\ \delta H/\delta\rho \\ \delta H/\delta A^{\beta}_{k} \\ \delta H/\delta{}^{*}E^{k}_{\beta} \\ \delta H/\delta G_\beta \end{array}\right| , \tag{2.33}$$

where s is any real constant. This Hamiltonian matrix reduces to (2.26a) when $^*\mathbf{E}$ is absent and $s=1$, and to (2.31) in the Abelian case. In comparison with (2.20a) for YM–MHD the Hamiltonian matrix b_3 in (2.33) for Yang–Mills charged fluids (YMCF) has been extended by adding a row and column for the dynamics of the variable $^*E^k_\beta$, the Yang–Mills analog of the electric displacement vector. The vector $^*E^k_\beta$ is dual to A^α_i in both the algebraic and metric sense: while A^α_i is a one-form taking values in the gauge algebra $\mathfrak{g}$, $^*E^k_\beta$ is an $(n-1)$ form taking values in the dual algebra $\mathfrak{g}^*$. The mathematical interpretation of the Poisson bracket determined from b_3 in (2.33) is given in Sec. III.

The relation of (2.33) for non-Abelian Yang–Mills charged fluids (YMCF) to the Hamiltonian matrix for CHD given in Gibbons, Holm, and Kupershmidt[27] is as follows. Let $\mathbf{M}$ define another momentum density via the map

$$M_i = P_i + \langle D^{**}_k E^k, A_i\rangle - \langle {}^*E^k, B_{ki}\rangle = P_i + ({}^*E^k_\alpha A^\alpha_i)_{,k} - {}^*E^k_\alpha A^\alpha_{k,i}, \tag{2.34}$$

while the other variables $(\rho,\mathbf{A},{}^*\mathbf{E},G)$ in (2.33) remain the same. The resulting Hamiltonian matrix in the new variables obtained via direct calculation using (2.25) is given (with $s=1$) by

$$\partial_t \begin{vmatrix} M_j \\ \rho \\ A^\beta_j \\ {}^*E^j_\beta \\ G_\beta \end{vmatrix} = - \begin{vmatrix} M_i\partial_j + \partial_i M_j & \rho\partial_j & 0 & 0 & [G_\alpha + {}^*(\mathrm{Div}\,\mathbf{E})_\alpha]\partial_j \\ \partial_i\rho & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \delta^\beta_\alpha\delta^i_j & (D_j)^\beta_\alpha \\ 0 & 0 & -\delta^\alpha_\beta\delta^i_j & 0 & t^\mu_{\beta\alpha}\ {}^*E^j_\mu \\ \partial_i[G_\beta + {}^*(\mathrm{Div}\,\mathbf{E})_\beta] & 0 & (D^*_i)^\alpha_\beta & t^\mu_{\beta\alpha}\ {}^*E^i_\mu & t^\mu_{\beta\alpha}G_\mu \end{vmatrix} \begin{vmatrix} \delta H/\delta M_i \\ \delta H/\delta\rho \\ \delta H/\delta A^\alpha_i \\ \delta H/\delta {}^*E^i_\alpha \\ \delta H/\delta G_\alpha \end{vmatrix}. \tag{2.35}$$

In the extended CHD Hamiltonian matrix (2.35) we denote [see (2.9) and (2.12)]

$$(D_j)^\beta_\alpha = \partial_j\delta^\beta_\alpha + t^\beta_{\alpha\mu}A^\mu_j\,, \tag{2.36a}$$

$$(D^*_i)^\alpha_\beta = \partial_i\delta^\alpha_\beta - t^\alpha_{\beta\mu}A^\mu_i\,, \tag{2.36b}$$

and [see (2.7) and (2.15)]

$$\begin{aligned} {}^*(\mathrm{Div}\,\mathbf{E})_\beta &= [\mathrm{Div}^*({}^*\mathbf{E})]_\beta = (D^*_k)^\alpha_\beta\ {}^*E^k_\alpha\,, \\ [\text{by (2.36b)}] &= \partial_k\ {}^*E^k_\beta - A^\mu_k\ {}^*E^k_\alpha t^\alpha_{\beta\mu}, \\ [\text{by (2.7)}] &= \{[\partial_k - \mathrm{ad}^*(A_k)]({}^*E^k)\}_\beta. \end{aligned} \tag{2.37}$$

The CHD Hamiltonian is (Gibbons, Holm, and Kupershmidt[27])

$$H = \int d^nx \left[\frac{1}{2\rho}|\mathbf{M} + G_\alpha\mathbf{A}^\alpha|^2 + U(\rho) + \frac{1}{2}\,{}^*\mathbf{E}_\alpha\cdot\mathbf{E}^\alpha + \frac{1}{4}\,{}^*B^{ik}_\alpha B^\alpha_{ik}\right], \tag{2.38}$$

with variational derivatives given by

$$\delta H = \int d^nx \left\{\left(-\frac{v^2}{2} + U'(\rho)\right)\delta\rho + (\mathbf{v}\cdot\mathbf{A}^\alpha)\delta G_\alpha + \mathbf{v}\cdot\boldsymbol{\delta}\mathbf{M} + \mathbf{E}^\alpha\cdot\boldsymbol{\delta}{}^*\mathbf{E}_\alpha + [G_\alpha v^i + {}^*(D_kB^{ki})_\alpha]\delta A^\alpha_i\right\}, \tag{2.39}$$

where we have integrated by parts and introduced the notation

$$\mathbf{v} = \rho^{-1}(\mathbf{M} + G_\alpha\mathbf{A}^\alpha), \tag{2.40}$$

$$\mathbf{E} = {}^*({}^*\mathbf{E}), \tag{2.41a}$$

$${}^*B_{ij} = {}^*(B_{ij}). \tag{2.41b}$$

The resulting Hamiltonian equations of motion for CHD are [using (2.22) to define the CHD Poisson bracket with Hamiltonian matrix b_4 in (2.35)]

$$\partial_t\rho = \{H,\rho\} = -\partial_i(\rho v^i), \tag{2.42a}$$

$$\partial_t A^\beta_j = \{H, A^\beta_j\} = -E^\beta_j - (D_j)^\beta_\alpha(\mathbf{v}\cdot\mathbf{A}^\alpha), \tag{2.42b}$$

$$\partial_t\ {}^*E^j_\beta = \{H,{}^*E^j_\beta\} = G_\beta v^j + {}^*(D_kB^{kj})_\beta - t^\mu_{\beta\alpha}\ {}^*E^j_\mu(\mathbf{v}\cdot\mathbf{A}^\alpha), \tag{2.42c}$$

$$\begin{aligned} \partial_t G_\beta = \{H,G_\beta\} = &-\partial_i([G_\beta + {}^*(\mathrm{Div}\,\mathbf{E})_\beta]v^i) \\ &- (D^*_i)^\alpha_\beta(G_\alpha v^i) - {}^*(D_iD_kB^{ki})_\beta \\ &- t^\mu_{\beta\alpha}\ {}^*E^i_\mu E^\alpha_i - t^\mu_{\beta\alpha}G_\mu(\mathbf{v}\cdot\mathbf{A}^\alpha). \end{aligned} \tag{2.42d}$$

Upon using (2.36b), (2.41a), and antisymmetry of B^{ki}, the G_β equation (2.42d) takes the form

$$\partial_t G_\beta = -\partial_i([G_\beta + {}^*(\mathrm{Div}\,\mathbf{E})_\beta]v^i) - \partial_i(G_\beta v^i). \tag{2.43}$$

This becomes simply the equation for gauge charge conservation upon setting

$$\mathrm{GAUSS} := G + {}^*(\mathrm{Div}\,\mathbf{E}) = 0, \tag{2.44}$$

and noting that this relation is preserved by the dynamics of (2.42b)–(2.42d), since

$$\begin{aligned} \partial_t(\mathrm{GAUSS}) &= -\mathrm{div}[(\mathrm{GAUSS})\mathbf{v}] + \mathrm{ad}^*(\mathbf{v}\cdot\mathbf{A})(\mathrm{GAUSS}) \\ &= -D^*_i[(\mathrm{GAUSS})v^i]. \end{aligned} \tag{2.45}$$

The proof of relation (2.45) is by direct computation, as follows. Using (2.44) we have

$$\begin{aligned} &\partial_t(\mathrm{GAUSS})_\beta \\ &\quad = \partial_t[G_\beta + (\partial_i\delta^\alpha_\beta - t^\alpha_{\beta\gamma}A^\gamma_i){}^*E^i_\alpha] \\ &\quad = \partial_t G_\beta - t^\alpha_{\beta\gamma}(\partial_t A^\gamma_i){}^*E^i_\alpha + (D^*_i)^\alpha_\beta\partial_t\ {}^*E^i_\alpha, \\ &\quad = -\partial_i[(\mathrm{GAUSS})_\beta v^i] - \partial_i(G_\beta v^i) \\ &\qquad + t^\alpha_{\beta\gamma}\ {}^*E^i_\alpha(D_i)^\gamma_\mu(\mathbf{v}\cdot\mathbf{A}^\mu) + (D^*_i)^\alpha_\beta(G_\alpha v^i) \\ &\qquad - (D^*_i)^\alpha_\beta[t^\mu_{\alpha\gamma}\ {}^*E^i_\mu(\mathbf{v}\cdot\mathbf{A}^\gamma)] \\ &\qquad [\text{by (2.42b)–(2.42d)}]. \end{aligned} \tag{2.46}$$

Thus, in the shorter notation of (2.7) and (2.11), and using (2.15), we have

$$\begin{aligned}
\partial_t(\mathrm{GAUSS}) &= \mathrm{Div}^*(G\mathbf{v}) - \partial_i[(\mathrm{GAUSS})v^i] - \partial_i(Gv^i) \\
&\quad + \mathrm{ad}^*(\mathbf{v}\cdot\mathbf{A})^*(\mathrm{Div}\,\mathbf{E}), \\
&= -\mathrm{div}[(\mathrm{GAUSS})\mathbf{v}] \\
&\quad + \mathrm{ad}^*(\mathbf{v}\cdot\mathbf{A})(\mathrm{GAUSS}) \quad [\text{by } (2.37\text{a})],
\end{aligned} \tag{2.47}$$

which recovers relation (2.45). Consequently, (2.43) becomes

$$\partial_t G_\beta = -\mathrm{div}(G_\beta \mathbf{v}), \tag{2.48}$$

upon using the nondynamical constraint (2.44), which may be regarded as an initial condition by virtue of (2.45).

Finally, we have the momentum equation

$$\begin{aligned}
\partial_t M_j = \{H,M_j\} &= -M_i\partial_j v^i - \partial_i(M_j v^i) \\
&\quad - \rho\partial_j[-v^2/2 + U'(\rho)].
\end{aligned} \tag{2.49}$$

Substituting (2.40) in the form

$$M_j = \rho v_j - \langle G,A_j\rangle, \tag{2.50}$$

into the momentum equation, (2.49), readily gives the velocity equation,

$$\begin{aligned}
&\rho[\partial_t v_j + v^i v_{j,i} + U'(\rho)_{,j}] \\
&\quad = -\partial_t\langle G,A_j\rangle - \partial_i\langle Gv^i,A_j\rangle - \langle G,A_i v^i_{,j}\rangle, \\
&\quad = \langle \partial_t G,A_j\rangle + \langle G,\partial_t A_j\rangle - \langle D_i^*(Gv^i),A_j\rangle \\
&\qquad - \langle Gv^i,D_iA_j\rangle \\
&\qquad - \langle G,(\mathbf{A}\cdot\mathbf{v})_{,j}\rangle - \langle G,v^iA_{i,j}\rangle \quad [\text{by } (2.12)], \\
&\quad = \langle G,\partial_t A_j + v^i(D_iA_j - D_jA_i) + (\mathbf{A}\cdot\mathbf{v})_{,j}\rangle \\
&\qquad [\text{by } (2.36) \text{ and } (2.48)], \\
&\quad = \langle G,-E_j - v^iB_{ij}\rangle \quad [\text{by } (2.42\text{b})].
\end{aligned} \tag{2.51}$$

Hence we recover precisely the motion equation for the fluid velocity in CHD. Namely, with $(\mathbf{v}\times\mathbf{B})_j := v^jB_{ij}$, in vector form,

$$\partial_t\mathbf{v} + (\mathbf{v}\cdot\nabla)\mathbf{v} = -\nabla U'(\rho) - \rho^{-1}\langle G,\mathbf{E} + \mathbf{v}\times\mathbf{B}\rangle. \tag{2.52}$$

This completes the derivation of the CHD equations (2.42a)–(2.42d) and (2.52) from the extended CHD Hamiltonian matrix in (2.35) and the CHD Hamiltonian H in (2.38). The physical interpretation of the CHD equations (2.42a)–(2.42d) and (2.52), and their derivation from kinetic theory is discussed in Gibbons, Holm, and Kupershmidt.[27]

The Yang–Mills "displacement vector" $^*\mathbf{E}$ has an interesting interpretation in spin-glass theory. Namely, $\mathbf{E} = \delta H/\delta^*\mathbf{E}$ is the disclination flux density, so that $^*\mathbf{E}$ is the disclination current density. More discussion of this interpretation is given in the concluding section.

III. MATHEMATICAL DISCUSSION

In this section we explain the general mathematical facts underlying the various Hamiltonian matrices appearing in the preceding section. This will supply the proof that the Jacobi identity is satisfied for all of the Poisson brackets in the preceding section.

A. General notation

Let $K = C^\infty(\mathbb{R}^n)$; $D = D(\mathbb{R}^n)$: Lie algebra of vector fields on $\mathbb{R}^n$; $\Lambda^k = \Lambda^k(\mathbb{R}^n)$: K module of differential k forms on $\mathbb{R}^n$; $X(\xi)$ denotes the Lie derivative of $\xi\in\Lambda^k$ with respect to $X\in D$; $\mathfrak{g}$: a finite-dimensional Lie algebra over $\mathbb{R}$; $\mathfrak{g}^*$: its dual; $(\ ,\ )$: a nondegenerate invariant symmetric bilinear form on $\mathfrak{g}$;

$$(e_1,\ldots,e_M)\text{: basis in } \mathfrak{g}, \text{ satisfying } [e_\alpha,e_\beta] = t^\gamma_{\alpha\beta}e_\gamma,$$

where $t^\gamma_{\alpha\beta}$ are the structure constants of $\mathfrak{g}$; $(e^1,\ldots,e^M)$, the dual basis in $\mathfrak{g}^*$; if $\sigma:\mathfrak{g}\to\mathrm{End}\,V$ is a representation of $\mathfrak{g}$, then $\sigma(a)(v)$ is denoted simply by $a.v$, for $a\in\mathfrak{g}$ and $v\in V$.

B. Lie algebra

We start with the Lie algebra $\mathfrak{g}_1$ (2.27). Its commutator is given by the formula

$$\begin{bmatrix} X^1 & X^2 \\ f^1\otimes a^1 & f^2\otimes a^2 \\ \omega^1\otimes b^1 & \omega^2\otimes b^2 \\ g^1 & , \quad g^2 \end{bmatrix} = \begin{matrix} [X^1,X^2] \\ X^1(f^2)\otimes a^2 - X^2(f^1)\otimes a^1 + f^1f^2\otimes[a^1,a^2] \\ X^1(\omega^2)\otimes b^2 - X^2(\omega^1)\otimes b^1 + f^1\omega^2\otimes a^1.b^2 - f^2\omega^1\otimes a^2.b^1 \\ X^1(g^2) - X^2(g^1) \end{matrix}\ , \tag{3.1}$$

where $X^i\in D$; $f^i,g^i\in K$; $\omega^i\in\Lambda^{n-1}$; $a^i\in\mathfrak{g}$; $b^i\in\mathfrak{g}^*$; $i = 1,2$; and, e.g., for $h\in\mathfrak{g}$ and pairing $\langle\ ,\ \rangle$ between $\mathfrak{g}^*$and $\mathfrak{g}$,

$$\langle a^1.b^2,h\rangle := -\langle b^2,[a^1,h]\rangle.$$

Claim: The commutator (3.1) defines a Lie algebra.

Proof: This results from the following general fact.

Theorem 3.1: Let Ω be a tensor field on $\mathbb{R}^n$, i.e., a K and D module, so that

$$X(f\omega) = fX(\omega) + X(f)\omega, \quad X\in D, \quad f\in K, \quad \omega\in\Omega. \tag{3.2}$$

Let $\sigma:\mathfrak{g}\to\mathrm{End}\,V$ be a representation of $\mathfrak{g}$. Then the following formula defines a Lie algebra $\overline{\mathfrak{g}}(\Omega,\sigma)$:

$$\begin{bmatrix} X^1 & X^2 \\ f^1\otimes a^1 & f^2\otimes a^2 \\ \omega^1\otimes v^1 & \omega^2\otimes v^2 \\ g^1 & , \quad g^2 \end{bmatrix} = \begin{matrix} [X^1,X^2] \\ X^1(f^2)\otimes a^2 - X^2(f^1)\otimes a^1 + f^1f^2\otimes[a^1,a^2] \\ X^1(\omega^2)\otimes v^2 - X^2(\omega^1)\otimes v^1 + f^1\omega^2\otimes a^1.v^2 - f^2\omega^1\otimes a^2.v^1 \\ X^1(g^2) - X^2(g^1) \end{matrix}\ , \tag{3.3}$$

where $X^i \in D$, f^i and $g^i \in K$, $\omega^i \in \Omega$, $a^i \in \mathfrak{g}$, $v^i \in V$, $i = 1,2$. A straightforward computation reduces the Jacobi identity for (3.3) to a set of identities of the form (3.2).

C. Generalized two-cocycles on the Lie algebra $\mathfrak{g}_1$

We now turn to the generalized two-cocycle on the Lie algebra $\mathfrak{g}_1$ responsible for the field-independent terms in the matrix (2.26a).

Proposition 3.2: The following formula defines a (generalized) two-cocycle ν_1 on $\mathfrak{g}_1 = \overline{\mathfrak{g}}(\Lambda^{n-1}, \mathrm{ad}^*)$:

$$\nu_1(1,2) = s(f^1\omega^2_{i,i}\langle b^2,a^1\rangle - f^2\omega^1_{i,i}\langle b^1,a^2\rangle), \quad s\in\mathbb{R}, \tag{3.4}$$

where $\langle\ ,\ \rangle$ is the natural pairing between $\mathfrak{g}^*$ and $\mathfrak{g}$; and the notation $\nu_1(1,2)$ is shorthand for

$$\nu_1\begin{pmatrix} X_1 & X_2 \\ \vdots\ , & \vdots \end{pmatrix}.$$

Recall (Kupershmidt,[32] Chap. viii) that a bilinear form ν on a Lie algebra $\mathfrak{g}'$ over K is called a (generalized) two-cocycle if

$$\nu(X,Y) \sim -\nu(Y,X), \quad X,Y\in\mathfrak{g}', \tag{3.5}$$

$$\nu([X,Y],Z) + \text{c.p.} \sim 0, \quad X,Y,Z\in\mathfrak{g}', \tag{3.6}$$

where "c.p." stands for "cyclic permutation"; and $a \sim b$ means $(a-b)\in\Sigma_i \operatorname{Im} \partial_i$, i.e., $(a-b)$ is a "divergence." One checks directly that ν_1 in (3.4) is indeed a two-cocycle on $\mathfrak{g}_1$.

D. Poisson bracket

The Poisson bracket associated to the two-cocycle ν_1 on the Lie algebra $\mathfrak{g}_1$ is computed by the standard rules of the general theory described in Kupershmidt,[32] Chap. viii (with n-dimensional volume element $d^n x$)

$$\begin{aligned}\{H,F\}_1 = -\int d^n x\Big\{&\frac{\delta F}{\delta P_i}\Big[(P_k\partial_i + \partial_k P_i)\Big(\frac{\delta H}{\delta P_k}\Big) + G_\beta\partial_i\Big(\frac{\delta H}{\delta G_\beta}\Big) + (\partial_k A^\beta_i - A^\beta_{k,i})\Big(\frac{\delta H}{\delta A^\beta_k}\Big) + \rho\partial_i\Big(\frac{\delta H}{\delta\rho}\Big)\Big]\\ &+\frac{\delta F}{\delta G_\alpha}\Big[\partial_k G_\alpha\Big(\frac{\delta H}{\delta P_k}\Big) + t^\gamma_{\alpha\beta}G_\gamma\frac{\delta H}{\delta G_\beta} + (-t^\beta_{\alpha\gamma}A^\gamma_k + \delta^\beta_\alpha s\partial_k)\Big(\frac{\delta H}{\delta A^\beta_k}\Big)\Big]\\ &+\frac{\delta F}{\delta A^\alpha_i}\Big[(A^\alpha_k\partial_i + A^\alpha_{i,k})\Big(\frac{\delta H}{\delta P_k}\Big) + (t^\alpha_{\beta\gamma}A^\gamma_i + \delta^\alpha_\beta s\partial_i)\Big(\frac{\delta H}{\delta G_\beta}\Big)\Big] + \frac{\delta F}{\delta\rho}\partial_k\rho\Big(\frac{\delta H}{\delta P_k}\Big)\Big\},\end{aligned} \tag{3.7}$$

where dual coordinates on $\mathfrak{g}_1^*$ are chosen to be

P_k dual to $\partial_k\in D$; G_α to $1\otimes e_\alpha\in K\otimes g^\alpha$; A_i to $(\partial_i \lrcorner\, d^n x)\otimes e^\alpha$; ρ to $1\in K$.

E. Spin-glass Hamiltonian matrix

The Hamiltonian matrix $\mathsf{b} = \mathsf{b}(\mathfrak{g}_1,\nu_1)$ associated to the Poisson bracket (3.7) via the standard rule

$$\{H,F\} \sim \frac{\delta F}{\delta u_i} b_{ij} \frac{\delta H}{\delta u_j}$$

is given by

$$\begin{array}{c|cccc|} & P_k & G_\beta & A^\beta_k & \rho \\ P_i & P_k\partial_i + \partial_k P_i & G_\beta\partial_i & \partial_k A^\beta_i - A^\beta_{k,i} & \rho\partial_i \\ G_\alpha & \partial_k G_\alpha & t^\gamma_{\alpha\beta}G_\gamma & -t^\beta_{\alpha\gamma}A^\gamma_k + \delta^\beta_\alpha s\partial_k & 0 \\ A^\alpha_i & A^\alpha_k\partial_i + A^\alpha_{i,k} & t^\alpha_{\beta\gamma}A^\gamma_i + \delta^\alpha_\beta s\partial_i & 0 & 0 \\ \rho & \partial_k\rho & 0 & 0 & 0 \end{array}\ . \tag{3.8}$$

This is b_1 in (2.26a) when $s = 1$.

F. Origin of the generalized two-cocycle

Since the two-cocycle (3.4) plays a crucial role in what follows, we explain its origin and unique features. Let $\widetilde{\Omega}$ be an additional tensor field on $\mathbb{R}^n$, and let $\vartheta: \Omega \to \widetilde{\Omega}$ be a homomorphism of D modules, i.e.,

$$\vartheta(X(\omega)) = X(\vartheta(\omega)), \quad X\in D, \quad \omega\in\Omega. \tag{3.9}$$

(For example, $\Omega = \Lambda^k$, $\widetilde{\Omega} = \Lambda^{k+1}$, $\vartheta = d$.) Then ϑ induces a natural Lie algebra homomorphism

$$\vartheta: \overline{\mathfrak{g}}(\Omega,\sigma) \to \overline{\mathfrak{g}}(\widetilde{\Omega},\sigma). \tag{3.10}$$

Therefore, from ϑ one obtains a Hamiltonian (i.e., canonical) map $\phi: C_{\bar{\mathfrak{g}}}(\Omega,\sigma) \to C_{\bar{\mathfrak{g}}}(\widetilde{\Omega},\sigma)$ on $C_{\bar{\mathfrak{g}}}$, the ring of functions on the dual to the Lie algebras $\bar{\mathfrak{g}}(\Omega,\sigma)$ and $\bar{\mathfrak{g}}(\tilde{\Omega},\sigma)$ [see Chap. viii (3.42) in Kupershmidt[32]]. In particular, take

$$\Omega = \Lambda^{n-1},\ \tilde{\Omega} = \Lambda^n,\ \vartheta = -d,\ V = \mathfrak{g}^*,\ \sigma = \mathrm{ad}^*,$$

and denote by η^α coordinates on $\mathfrak{g}_2^* = \bar{\mathfrak{g}}(\Lambda^n,\mathrm{ad}^*)^*$ dual to $d^n x \otimes e^\alpha$. Then by formula viii (3.42) in Kupershmidt,[32] the map ϕ^* can be written in the form

$$P_i = P_i;\quad G_\alpha = G_\alpha;\quad A_i^\alpha = \eta^\alpha_{,i};\quad \rho = \rho; \tag{3.11}$$

and ϕ^* is a Hamiltonian map between the cocycles Poisson bracket $(3.7)|_{s=0}$ and the Poisson bracket on $\mathfrak{g}_2^* = \bar{\mathfrak{g}}(\Lambda^n,\mathrm{ad}^*)^*$,

$$\begin{aligned}\{H,F\} = &-\int d^n x \left\{\frac{\delta F}{\delta P_i}\left[(P_k\partial_i + \partial_k P_i)\left(\frac{\delta H}{\delta P_k}\right) + G_\beta \partial_i\left(\frac{\delta H}{\delta G_\beta}\right) - \eta^\beta_{,i}\frac{\delta H}{\delta \eta^\beta} + \rho\partial_i\left(\frac{\delta H}{\delta\rho}\right)\right]\right.\\ &\left. + \frac{\delta F}{\delta G_\alpha}\left[\partial_k G_\alpha\left(\frac{\delta H}{\delta P_k}\right) + t^\gamma_{\alpha\beta} G_\gamma \frac{\delta H}{\delta G_\beta} + t^\beta_{\alpha\gamma}\eta^\gamma\frac{\delta H}{\delta\eta^\beta}\right] + \frac{\delta F}{\delta\eta^\alpha}\left(\eta^\alpha_{,k}\frac{\delta H}{\delta P_k} - t^\alpha_{\beta\gamma}\eta^\gamma\frac{\delta H}{\delta G_\beta}\right) + \frac{\delta F}{\delta\rho}\partial_k\rho\left(\frac{\delta H}{\delta P_k}\right)\right\}.\end{aligned} \tag{3.12}$$

Now, as a K module, $\mathfrak{g}_2 = \bar{\mathfrak{g}}(\Lambda^n, \mathrm{ad}^*)$, has inside it two submodules: $K \otimes \mathfrak{g}$ and $\Lambda^n \otimes \mathfrak{g}^*$, which are mutually dual as D modules, i.e.,

$$\langle X(\bar{b}),\bar{a}\rangle + \langle \bar{b},X(\bar{a})\rangle \sim 0,\quad \bar{a}\in K\otimes\mathfrak{g},\quad \bar{b}\in\Lambda^n\otimes\mathfrak{g}^*. \tag{3.13}$$

This means that we may have a symplectic two-cocycle ν_2 on $\mathfrak{g}_2$

$$\nu_2(1,2) = -s(f^1\psi^2\langle b^2,a^1\rangle - f^2\psi^1\langle b^1,a^2\rangle),\quad s\in\mathbb{R},\quad \psi^i\in\Lambda^n. \tag{3.14}$$

And indeed, ν_2 is a two-cocycle on $\mathfrak{g}_2$, as one verifies by a direct computation. [A verification is required since, for non-Abelian $\mathfrak{g}$, $K\otimes\mathfrak{g}$ acts nontrivially on $\Lambda^n\otimes\mathfrak{g}^*$; otherwise (3.13) would have guaranteed that (3.14) is a two-cocycle.] Now, since the map ϕ^* in (3.11) is constant coefficient, it transforms two-cocycles on $\mathfrak{g}_2$ into two-cocycles on $\mathfrak{g}_1$; in particular, ν_2 is transformed into ν_1. From this discussion, one concludes that if $\Omega\otimes V \neq \Lambda^{n-1}\otimes\mathfrak{g}^*$, then one cannot have a two-cocycle on $\bar{\mathfrak{g}}(\Omega,\sigma)$ similar to ν_1, since a symplectic two-cocycle of the type ν_1 exists only on $\bar{\mathfrak{g}}(\Lambda^n, \mathrm{ad}^*)$. This observation saves us from a futile search for new two-cocycles in the extended YMCF case, when *E variables (dual to $\Lambda^1\otimes\mathfrak{g}$) come into the picture.

G. Lie algebra $\mathfrak{g}_3$

The Lie algebra $\mathfrak{g}_3 = \bar{\mathfrak{g}}(\Lambda^{n-1}\oplus\Lambda^1, \mathrm{ad}^*\oplus\mathrm{ad})$ (*E is included), has commutator (cf. Theorem 3.1)

$$\begin{bmatrix} X^1 & X^2 \\ f^1\otimes a^1 & f^2\otimes a^2 \\ \omega^1\otimes b^1 & \omega^2\otimes b^2 \\ \mu^1\otimes\tilde{a}^1 & \mu^2\otimes\tilde{a}^2 \\ g^1 \quad , & g^2 \end{bmatrix} = \begin{matrix} [X^1,X^2] \\ X^1(f^2)\otimes a^2 - X^2(f^1)\otimes a^1 + f^1 f^2\otimes[a^1,a^2] \\ X^1(\omega^2)\otimes b^2 - X^2(\omega^1)\otimes b^1 + f^1\omega^2\otimes a^1.b^2 - f^2\omega^1\otimes a^2.b^1 \\ X^1(\mu^2)\otimes\tilde{a}^2 - X^2(\mu^1)\otimes\tilde{a}^1 + f^1\mu^2\otimes[a^1,\tilde{a}^2] - f^2\mu^1\otimes[a^2,\tilde{a}^1] \\ X^1(g^2) - X^2(g^1)\end{matrix}\quad , \tag{3.15}$$

where $X^i\in D, f^i,g^i\in K, \omega^i\in\Lambda^{n-1}, \mu^i\in\Lambda^1; a^i,\tilde{a}^i\in\mathfrak{g}; b^i\in\mathfrak{g}^*; i = 1,2.$

H. Remarks

(a) $\mathfrak{g}_3$ contains $\mathfrak{g}_1$ as a subalgebra, and $\mathfrak{g}_3$ itself is a semidirect product of $\mathfrak{g}_1$ and $\Lambda^1\otimes\mathfrak{g}$. Hence, there is a two-cocycle $\tilde{\nu}_1$ on $\mathfrak{g}_3$, which coincides with ν_1 on $\mathfrak{g}_1$ and vanishes when one of its arguments belongs to $\Lambda^1\otimes\mathfrak{g}$:

$$\tilde{\nu}_1(1,2) = s(f^1\omega^2_{i,i}\langle b^2,a^1\rangle - f^2\omega^1_{i,i}\langle b^1,a^2\rangle),\quad s\in\mathbb{R}. \tag{3.16}$$

(b) There is also a new symplectic two-cocycle on $\mathfrak{g}_3$,

$$\nu_3(1,2) = c_1(\omega^1\wedge\mu^2\langle b^1,\tilde{a}^2\rangle - \omega^2\wedge\mu^1\langle b^2,\tilde{a}^1\rangle),\quad c_1\in\mathbb{R}. \tag{3.17}$$

(c) The new Poisson bracket associated to the two-cocycle $\tilde{\nu}_1 + \nu_3$ on $\mathfrak{g}_3$ equals

$$\begin{aligned}\{H,F\}_3 = \{H,F\}_1 &+ \int d^n x \left[\frac{\delta F}{\delta P_i}({}^*E^k_\beta\partial_i - \partial_j\,{}^*E^j_\beta\delta^i_k) + \frac{\delta F}{\delta G_\alpha}t^\gamma_{\alpha\beta}\,{}^*E^k_\gamma + \frac{\delta F}{\delta A^\alpha_i}c_1\delta^\alpha_\beta\delta^k_i\right]\left(\frac{\delta H}{\delta\,{}^*E^k_\beta}\right)\\ &+ \frac{\delta F}{\delta\,{}^*E^i_\alpha}\left[(\partial_k\,{}^*E^i_\alpha - {}^*E^j_\alpha\partial_j\delta^i_k)\left(\frac{\delta H}{\delta P_k}\right) + t^\gamma_{\alpha\beta}\,{}^*E^i_\gamma\frac{\delta H}{\delta G_\beta} - c_1\delta^\beta_\alpha\delta^i_k\frac{\delta H}{\delta A^\beta_k}\right],\end{aligned} \tag{3.18}$$

where $\{H,F\}_1$ is the Poisson bracket corresponding to $\mathfrak{b}_1$ in (3.8), and ${}^*E^k_\beta$ is dual to $dx^k\otimes e_\beta$, in both the metric, and the Lie-algebraic senses. For $c_1 = 1$, the Hamiltonian matrix associated to (3.18) is given in (2.33).

I. Lie algebra $\mathfrak{g}_4$

Let $\mathfrak{g}_4 = \bar{\mathfrak{g}}(\Lambda^{n-2}, \mathrm{ad}^*)$. This is the Lie algebra with the commutator (3.3) for $\omega^i\in\Lambda^{n-2}$ and $v^i\in\mathfrak{g}^*$, $i = 1,2$. The corresponding Hamiltonian matrix is given by formula (2.28), provided one lets B^α_{ij} be the coordinate dual to $(\partial_i \lrcorner\, \partial_j \lrcorner\, d^n x)\otimes e^\alpha\in\Lambda^{n-2}\otimes\mathfrak{g}^*$.

IV. CONCLUSIONS

We have considered the analogy between spin glasses and Yang–Mills fluids (CHD) within the Hamiltonian framework. Our results complete this analogy, according to the following "dictionary."

spin glass		Yang–Mills fluid	
ρ,	defect inertial-mass density	ρ,	mass density of fluid carrying gauge charge
$\mathbf{v}$,	fluid velocity	$\mathbf{v}$,	fluid velocity
$\mathbf{K}$,	hydrodynamic momentum density of defects	$\mathbf{M}$,	total momentum density, including YM field momentum
B_{ij},	disclination density	B_{ij},	Yang–Mills magnetic field
F_{i0},	disclination current density	$^*E^i$,	Yang–Mills electric displacement vector

Along the way, we have noticed an interesting phenomenon in YM–MHD and CHD, namely, the existence of two different Poisson brackets for the non-Abelian case and a one-parameter family of Poisson brackets for the Abelian case, in the **A** representation for CHD, see Eq. (2.33).

Physically, our conclusion is that the analogy between spin-glass theory and Yang–Mills charged fluids is very close, on the level of the Hamiltonian formalism. Specifically, the Hamiltonian matrices are identical for the Volovik–Dotsenko spin-glass theory and Yang–Mills MHD. In addition, the Hamiltonian matrix (2.33) in the Yang–Mills charged-fluid representation provides a potentially interesting extension of the Volovik–Dotsenko spin-glass theory, by providing a dynamical equation for the disclination current density $^*\mathbf{E}$, which is the spin-glass analog of the Yang–Mills electric displacement vector.

Our basic mathematical observations are these: the highly nonlinear candidate (2.20) for the Poisson bracket in Volovik and Dotsenko,[24] when transformed to appropriate (natural) variables, becomes of affine type and is thus associated to a certain Lie algebra, called $\mathfrak{g}_1$, and a two-cocycle, called ν_1, on $\mathfrak{g}_1$. It turns out that $\mathfrak{g}_1$ is a subalgebra of another Lie algebra, $\mathfrak{g}_3$, which closely resembles the chromohydrodynamics Lie algebra $\mathfrak{g}_2$. The Lie algebra $\mathfrak{g}_2$ is, in turn, another subalgebra of $\mathfrak{g}_3$. Moreover, the two-cocycle ν_1 on g_1 is a restriction on $\mathfrak{g}_1 \subset \mathfrak{g}_3$ of a certain two-cocycle $\tilde{\nu}_1$ on $\mathfrak{g}_3$. Furthermore, there is another, canonical, two-cocycle ν_3 on $\mathfrak{g}_3$, whose restriction on $\mathfrak{g}_1$ vanishes and whose restriction on $\mathfrak{g}_2$ produces precisely the canonical $^*\mathbf{E}$-$\mathbf{A}$ structure in CHD.

Roughly speaking, the absence of a dynamical equation for $^*\mathbf{E}$ in Volovik and Dotsenko[24] is of the same nature as the absence of displacement current. The dynamical equation for $^*\mathbf{E}$ is present only in the full electromagnetic or Yang–Mills field equations, or in an extended theory of spin-glass dynamics accounting for time dependence of the disclination current density, F_{i0}. In that case, the present theory would provide the dynamics by using Poisson bracket (2.33), in conjunction with an appropriate choice for the Hamiltonian.

[1] B. I. Halperin and W. M. Saslow, Phys. Rev. B **16**, 2154 (1977).

[2] A. F. Andreev, Zh. Eksp. Teor. Fiz. **74**, 786 (1978) [Sov. Phys. JETP **47**, 411 (1978)].

[3] G. Toulouse, Phys. Rep. **49**, 267 (1979).

[4] C. L. Henley, H. Sompolinsky, and B. I. Halperin, Phys. Rev. B **25**, 5849 (1982).

[5] A. J. Bray and M. A. Moore, J. Phys. C **15**, 2417 (1982).

[6] W. M. Saslow, Phys. Rev. B **27**, 6873 (1983).

[7] C. L. Henley, Ann. Phys. (NY) **156**, 368 (1984).

[8] G. Toulouse and M. Kléman, J. Phys. Lett. 37, L-149 (1976).

[9] G. E. Volovik and V. P. Mineev, Zh. Eksp. Teor. Fiz. Pisma **23**, 647 (1976) [JETP Lett. **23**, 593 (1976)].

[10] N. D. Mermin, Rev. Mod. Phys. **51**, 591 (1979).

[11] L. Michel, Rev. Mod. Phys. **52**, 617 (1980).

[12] I. E. Dzyaloshinskii and G. E. Volovik, J. Phys. (Paris) **39**, 693 (1978).

[13] I. E. Dzyaloshinskii and G. E. Volovik, Ann. Phys. (NY) **125**, 67 (1980).

[14] J. A. Hertz, Phys. Rev. B **18**, 4875 (1978).

[15] J. José and J. A. Hertz, Bull. Am. Phys. Soc. **24**, 304 (1979).

[16] I. E. Dzyaloshinskii, in *Lecture Notes in Physics*, Vol. 115 (Springer, Berlin, 1980), p. 204.

[17] I. E. Dzyaloshinskii, Zh. Eksp. Teor. Fiz. Pisma **37**, 190 (1983) [JETP Lett. **37**, 227 (1983)].

[18] K. H. Fischer, Phys. Status Solidi B **116**, 357 (1983); **130**, 13 (1985).

[19] S. S. Rozhkov, Phys. Lett. A **106**, 309 (1984).

[20] K. Kawasaki and H. R. Brand, Ann. Phys. (NY) **160**, 420 (1985).

[21] Y. Bouligand, B. Derrida, V. Poénaru, Y. Pomeau, and G. Toulouse, J. Phys. (Paris) **39**, 863 (1978).

[22] W. M. Saslow, Phys. Rev. B **22**, 1174 (1980).

[23] D. Chowdhury and A. Mookerjee, Phys. Rep. **114**, 1 (1984).

[24] G. E. Volovik and V. S. Dotsenko, Zh. Eksp. Teor. Fiz. **78**, 132 (1980) [Sov. Phys. JETP **51**, 65 (1980)].

[25] I. M. Khalatnikov and V. V. Lebedev, J. Low Temp. Phys. **32**, 789 (1978); Prog. Theor. Phys. Suppl. **69**, 269 (1980).

[26] D. D. Holm and B. A. Kupershmidt, Phys. Lett. A **91**, 425 (1982).

[27] J. Gibbons, D. D. Holm, and B. A. Kupershmidt, Phys. Lett. A **90**, 281 (1982); Physica D **6**, 179 (1983).

[28] D. D. Holm and B. A. Kupershmidt, Phys. Rev. D **30**, 2557 (1984); Phys. Lett. A **105**, 225 (1984).

[29] D. D. Holm, J. E. Marsden, T. Ratiu, and A. Weinstein, Phys. Rep. **123**, 1 (1985).

[30] G. E. Volovik and I. E. Dzyaloshinskii, Zh. Eksp. Teor. Fiz. **75**, 1102 (1978) [Sov. Phys. JETP **48**, 555 (1979)].

[31] W. Drechsler and M. E. Meyer, "Fiber bundle techniques in gauge theories," *Lecture Notes in Physics*, Vol. 67 (Springer, Berlin, 1977).

[32] B. A. Kupershmidt, *Discrete Lax Equations and Differential-Difference Calculus* (Asterique, Paris, 1985).

[33] D. D. Holm, Phys. Lett. A **114**, 137 (1986).

[34] J. P. Friedberg, Rev. Mod. Phys. **54**, 801 (1982).

Global hyperbolicity of a spatially closed space-time

Yoshihiro Matori
Department of Mathematics, Faculty of Science, Kyushu University, Fukuoka, 812, Japan

(Received 4 June 1987; accepted for publication 9 September 1987)

A spatially closed space-time is shown to be globally hyberbolic.

I. INTRODUCTION

Global hyperbolicity is one of the most important causality conditions on a space-time. For instance, a globally hyperbolic space-time has nice properties[1] such as the finiteness and the continuity of Lorentzian distance function, and the existence of a maximal geodesic segment between a pair of causally related points. Furthermore, Geroch[2] proved that a globally hyperbolic space-time (M,g) of dimension $n+1$ is homeomorphic to $S\times\mathbb{R}$ with S a manifold of dimension n. He also pointed out that the converse of this is not necessarily true. In this paper, we shall show that the converse is true if S is compact and $S\times\{a\}$ is locally acausal for every $a\in\mathbb{R}$. This implies that de Sitter space-time, Einstein's static universe, and spatially closed Robertson–Walker space-time are all globally hyperbolic.

For our notation and conventions we mostly follow Hawking–Ellis.[1] In particular, by a space-time (M,g) we mean a connected time-oriented C^∞ Lorentzian manifold of dimension $n+1$ with C^∞ Lorentzian metric g. A space-time (M,g) is called globally hyperbolic if (M,g) is strongly causal and $J^+(p)\cap J^-(q)$ is compact for every $p,q\in M$, where $J^+(p)$ [resp. $J^-(p)$] is the causal future (resp. past) of p. A subset A of M is called acausal if every nonspacelike curve intersects A at most once. A subset B of M is called locally acausal if every point of B has a neighborhood in which B is acausal.

II. MAIN RESULT

Theorem: Let (M,g) be a space-time homeomorphic to $S\times\mathbb{R}$, where S is a manifold of dimension n. If S is compact and $S\times\{a\}$ is locally acausal for every $a\in\mathbb{R}$, then (M,g) is globally hyperbolic.

Proof: Recall that a space-time is globally hyperbolic iff it has a Cauchy surface, i.e., a boundaryless imbedded submanifold which every inextendible nonspacelike curve intersects exactly once.[1,2] Hence we are going to show that $S\times\{a\}$ is a Cauchy surface for every $a\in\mathbb{R}$.

Let γ be an inextendible nonspacelike curve in M and $p\in\gamma$. Note that the $\mathbb{R}$ coordinate can be regarded as a time function, which we denote by t [so each $S\times\{a\}$ is (globally) acausal]. In case $t(p)>a$, $\gamma\cap J^-(p)$ is a past-inextendible nonspacelike curve and is contained in the region $S\times(-\infty,t(p)]$. Then it follows from the hypotheses of the theorem that (M,g) is strongly causal.[3] Hence $\gamma\cap J^-(p)$ cannot be imprisoned[4] in the compact set $S\times[a,t(p)]$. This means that $\gamma\cap J^-(p)$ enters into the region $S\times(-\infty,a]$. Since γ and t are continuous, $\gamma\cap J^-(p)$ intersects $S\times\{a\}$.

In case $t(p)<a$, the similar discussion shows that $\gamma\cap J^+(p)$ intersects $S\times\{a\}$. Hence every inextendible nonspacelike curve intersects $S\times\{a\}$. Since it has been noted that $S\times\{a\}$ is acausal, it follows that $S\times\{a\}$ is a Cauchy surface, and hence (M,g) is globally hyperbolic.

Remarks: (1) The local acausality of $S\times\{a\}$ for every $a\in\mathbb{R}$ is necessary. For example, let (M,g) be a space-time homeomorphic to $S^1\times\mathbb{R}$ and

$$g=(\cosh t-1)^2(-dt^2+dx^2)-dt\,dx,\quad (x,t)\in S^1\times\mathbb{R},$$

where S^1 is the one-dimensional sphere. This space-time has a closed null geodesic $S^1\times\{0\}=t^{-1}(0)$. Thus (M,g) is not globally hyperbolic.

(2) The compactness of S is also necessary. For example, the universal covering manifold of anti-de Sitter space-time is not globally hyperbolic,[1,2] but is homeomorphic to $\mathbb{R}^n\times\mathbb{R}$ and $\mathbb{R}^n\times\{a\}$ is acausal for every $a\in\mathbb{R}$.

[1] S. W. Hawking and G. F. R. Ellis, *The Large Scale Structure of Space-time* (Cambridge U. P., Cambridge, 1973).

[2] R. Geroch, J. Math. Phys. **11**, 437 (1970).

[3] See Theorem 2.1. of H. J. Seifert, Gen. Relativ. Gravit. **8**, 815 (1977).

[4] See Proposition 6.4.7. of Ref. 1.

0022-2488/88/010031-01$02.50

A solution to the one-dimensional missing moment problem

Carlos R. Handy
Physics Department, Atlanta University, Atlanta, Georgia 30314

(Received 7 November 1986; accepted for publication 26 August 1987)

The one-dimensional "missing moment problem" is solved using Padé analysis. The realization of this affords the most efficient framework within which to apply a Hankel–Hadamard analysis for generating rapidly convergent bounds to quantum eigenvalues. The method is applied to the quartic potential problem.

I. INTRODUCTION

In a recent work, Handy and Bessis[1] discovered that the implementation of a "moments problem"[2] reformulation of the one-dimensional Schrödinger equation, for arbitrary rational fraction potential, could yield exponentially convergent lower and upper bounds to the quantum eigenvalues. This is an important endeavor because it is the best way of assessing the accuracy of one's eigenenergy estimate. These concerns are particularly relevant for singular perturbation–strongly coupled systems not amenable to conventional perturbation analysis. As argued in Ref. 1, a moments approach is ideally suited for these kinds of problems.

In this work we focus on an important aspect of the above moments approach. Because of the unprecedented nature of the theory developed by Handy and Bessis, we briefly outline here the pertinent issues, as well as clarify the contribution made by the present work. A more detailed discussion is presented in the following section.

Given the Schrödinger equation for some arbitrary rational fraction potential, it is straightforward to define a moments equation. This will generally be a recursive relation by which the Hamburger moments of the wave function, $\mu_p = \int dx\, x^p \Psi$, can be obtained once the energy E and a certain finite number of initial moments, $\mu_1,\dots,\mu_m$, are specified. Thus we may express this by $\mu_i = F_i[E,\mu_1,\dots,\mu_m]$, for $1 \leqslant i \leqslant \infty$. Because the ground state wave function is non-negative[1] one can then use the Hankel–Hadamard[2] inequalities to define a hierarchy of constraints, $\Delta_{k,l}(E,\mu_1,\dots,\mu_m) > 0$, for E and the m-missing moments (one can always normalize things to $\mu_0 = 1$). The results of Handy and Bessis[1] show that the Hankel–Hadamard inequalities are sufficiently strong to yield exponentially convergent lower and upper bounds to E and the missing moments. In this way, excellent accurate physical values are realized for all of these quantities. The above results are readily extendable to excited states. For simplicity, we limit the present discussion to the ground state only.

Clearly, the Hankel–Hadamard inequalities define a succession of rapidly decreasing $(m + 1)$-dimensional subregions within the $(m + 1)$-tuple space defined by E and the m-missing moments. In practice, for systems with too many missing moments $(m > 2)$, the identification of these subregions becomes costly. It is therefore clear that the identification of some formalism by which the number of missing moments can be reduced or completely eliminated is an important concern. *The basic contribution of this work is the attainment of one particular formalism that can systematically reduce the number of missing moments to zero!* The manner in which this is done highlights some very profound and very general characteristics of the moment method.

Before proceeding to the next section, it is important to make more precise the claim made above. As was shown in Ref. 1, for some systems, such as the sextic problem, $-\Psi'' + (mx^2 + x^6)\Psi = E\Psi$, one can find a special wave function representation space in which the non-negativity property of the ground state solution is preserved. The new representation is defined by the zeroth-order WKB expression $\Phi(x) = \exp(-x^4/4)\Psi(x)$. In addition to preservation of non-negativity, the new representation also leaves unphysical Ψ-space solutions as unphysical Φ-space solutions. A solution is unphysical if its moments are infinite. It is immediate to show that in the Φ representation space, there are no missing moments! The original sextic problem has two missing moments. Thus for this case there exists a *global* transformation that completely eliminates all the missing moments. For the quartic potential problem, $-\Psi'' + (mx^2 + x^4)\Psi = E\Psi$, there is no global transformation to eliminate the one-missing moment nature of this system.[1] Despite this, it is still possible to find wave function representation spaces where the first Q moments depend upon E only, while the remaining moments $(\mu_{Q+1},\dots)$ depend upon E and the missing moments. This Q is arbitrary. Thus this formulation yields an effective zero-missing moment problem up to order Q. Because the Hankel–Hadamard inequalities yield exponentially convergent bounds, in practical terms, this alternate formulation for reducing the missing moment problem is just as effective as finding a global transformation to completely remove the missing moments. *This is the contribution of the present work.* We demonstrate our formalism by applying it to the quartic potential problem with $m = 0$.

II. A SHORT REVIEW

The central theme of the work of Handy and Bessis[1] is that because the bosonic ground state wave function is non-negative, one may use the moments problem[2] to quantize the system. It is also possible to extend this formulation to excited states for which the wave function is non-negative.[3] Consider the sextic potential problem

$$-\Psi'' + (mx^2 + gx^6) = E\Psi. \tag{1}$$

0022-2488/88/010032-04$02.50

The pth-order Hamburger moment is defined by

$$\mu(p) = \int_{-\infty}^{+\infty} dx\, x^p \Psi(x).$$

The ground state wave function is non-negative ($\Psi \geqslant 0$), symmetric, and asymptotically fast decreasing[1] so that the moments exist. Through an immediate integration by parts Eq. (1) yields a moments recursion relation:

$$-p(p-1)\mu(p-2) + m\mu(p+2) + g\mu(p+6) = E\mu(p). \qquad (2)$$

The symmetric nature of the implicit ground state solution allows us to also use a Stieltjes moment representation which, although not necessary, is more convenient $[\mu(p=\text{odd}) \equiv 0]$. Specifically, through a change of variables ($x^2 \equiv y$), the even-order Hamburger moments are equivalent to the Stieltjes moments $\hat{\mu}$ of a modified function measure, $f(y) = \Psi/\sqrt{y}$:

$$\mu(2p) = \int_0^{+\infty} dy\, y^p\, \Psi \frac{(\sqrt{y})}{\sqrt{y}} \equiv \hat{\mu}(p), \quad \text{a Stieltjes moment.} \qquad (3)$$

The moments problem[2] concerns the specification of the conditions under which the moments may be used to prove that a function measure is non-negative. The Stieltjes moment problem was first formulated in 1895[2]; followed by the Hamburger formulation in 1920.[2] Each of these concerns different types of moments and function domains, as suggested by the definitions for $\hat{\mu}$ and μ. In principle, either formulation can be used for quantizing our physical system.

Let $\Delta(m,n)$ denote a particular Hankel–Hadamard (HH) determinant, as defined by

$$\Delta(m,n) = \begin{vmatrix} \mu(m) & \mu(m+1) & \cdots & \mu(m+n) \\ \vdots & & & \vdots \\ \mu(m+n) & & \cdots & \mu(m+2n) \end{vmatrix}. \qquad (4)$$

The corresponding HH determinant for the Stieltjes case will be denoted by $\hat{\Delta}(m,n)$. The Stieltjes moment theorem states[4] that the necessary and sufficient conditions for $f(y)$ to be non-negative throughout the interval $[0,\infty]$ are

$$\hat{\Delta}(0,n) > 0 \quad \text{and} \quad \hat{\Delta}(1,n) > 0, \quad \text{for all } n. \qquad (5)$$

The Hamburger moment theorem states[4] that the necessary and sufficient conditions for $\Psi(x)$ to be non-negative throughout the interval $(-\infty,\infty)$ are

$$\Delta(0,n) > 0, \quad \text{for all } n. \qquad (6)$$

At an intuitive level, Eq. (5) follows from Eq. (6), if we use Eq. (6) on $\psi(x)$ and $x\psi(x)$.

Bearing in mind Eq. (3), a recursion relation for the Stieltjes moments follows from Eq. (2):

$$\hat{\mu}(p+3) = (1/g)[E\hat{\mu}(p) - m\hat{\mu}(p+1) + 2p(2p-1)\hat{\mu}(p-1)], \quad \text{for } p \geqslant 0. \qquad (7)$$

Because $\Psi(x)$ has an arbitrary normalization and is non-negative, we may set $\hat{\mu}(0) \equiv 1$. In addition, it is also clear that all of the moments are dependent on E, $\hat{\mu}(1)$ and $\hat{\mu}(2)$. Thus this is a two-missing moment problem. Nevertheless, as alluded to in the Introduction, the HH inequalities in Eq. (5) rapidly define a very small three-dimensional subregion centered around the physical values for E, $\hat{\mu}(1)$, and $\hat{\mu}(2)$. Thus, through Eq. (7), the HH determinants acquire an implicit dependence of the type

$$\hat{\Delta}(0,n;E,\hat{\mu}_1,\hat{\mu}_2) > 0, \quad \hat{\Delta}(1,n;E,\hat{\mu}_1,\hat{\mu}_2) > 0, \quad \text{for } n \geqslant 0. \qquad (8)$$

On a computer, working with at most Q Stieltjes moments $[\hat{\mu}(p),\, p \leqslant Q]$, a three-dimensional partitioning of a given region is defined. At each grid point, the corresponding HH inequalities are tested. In this manner, a consistent subregion can be found. Using $Q \leqslant 12$, for the sextic we have $\hat{\mu}_1 = 0.47$, $\hat{\mu}_2 = 0.56$, and $E = 1.436$.

Clearly the above program can be impractical on slow computers. One would prefer to transform the problem into another with fewer, or no, missing moments. For the sextic case the latter is possible. Indeed, by using

$$\phi(x) = \exp(S(x))\Psi(x), \qquad (9)$$

where $S(x) = -\frac{1}{4}\sqrt{g}x^4$, the zeroth-order WKB term, it is found that a Stieltjes moment ϕ-space analysis is of zero-missing moment type.[1]

The closed form of Eq. (9) does not always work in reducing the number of missing moments. A specific example of interest to us is the quartic potential problem, $V(x) = mx^2 + x^4$. It is a one-missing moment problem that cannot be simplified through closed expressions of the above type. We will return to this shortly.

It is important to realize that Eq. (9) encompasses some very important and profound generalities provided by our moments perspective. Specifically, let us categorize the key ingredients of our overall program: (I) work within representation spaces where the ground state wave function is non-negative; (II) work within a representation space where the physical moments are finite; (III) work within a representation space where the moments can be readily solved for (preferably in terms of a recursion relation); and (IV) choose representation with smallest number of missing moments. In general, if $T(x) > 0$, then $\phi(x) \equiv T(x)\Psi(x)$ defines a suitable representation, with respect to condition (I). Condition (II) is important because it focuses on the asymptotic behavior of the desired solution. Insofar as the asymptotic behavior of solutions to Schrödinger's equation are governed by zeroth WKB analysis, it is clear that the latter is an important concern to our overall program. Condition (III) is the least necessary, as a matter of principle. It is clearly the most convenient. Condition (IV) is self-evident.

There are many representations satisfying (I)–(III), as will become evident in the next section. If we insist on elegant, closed transformations that completely eliminate the missing moments, then there is little likelihood of finding them, except for various special cases.

From a practical standpoint, bearing in mind the fact that only a small number of moments are actually used,[1,3] it is clear that the attainment of a complete zero-missing moment representation is unnecessary. This realization leads to our principal contribution, developed in the next section.

III. SOLVING THE ONE-DIMENSIONAL MISSING MOMENT PROBLEM

Given an n-missing moment system,

$$\hat{\mu}(p+n+1) = F(p,E,\hat{\mu}_1,\hat{\mu}_2,...,\hat{\mu}_n), \quad p \geqslant 0, \tag{10}$$

a ϕ representation can be found for which the first $Q+1$ moments $[\tilde{\mu}(0),...,\tilde{\mu}(Q)]$ depend on only $(n-1)$-missing moments. The remaining moments $\{\tilde{\mu}(Q+1),...\}$ will depend on n-missing moments. The specification of Q is arbitrary.

The inductive application of the above can be used to convert an n-missing moment system into one for which the first $T+1$ moments depend on *no* missing moments.

The proof of the above is presented in the context of the sextic potential discussed in the preceding section. A numerical example is given in the following section.

Consider the sextic problem $-\Psi'' + (mx^2 + gx^6)\Psi = E\Psi$. Define the polynomial transformation

$$\phi(x) = \left|\sum_{i=0}^{I} C_i x^{2i}\right|^2 \Psi(x). \tag{11}$$

The C_i's are complex coefficients. In terms of the Ψ-Stieltjes moments $\{\hat{\mu}\}$, the ϕ-Stieltjes moments $\{\tilde{\mu}\}$ become

$$\tilde{\mu}(p) = \sum_{i,j=0}^{I} C_i^* C_j \hat{\mu}(p+i+j). \tag{12}$$

The two-missing moment nature of the sextic potential, as well as the linear homogeneous nature of the difference equation in Eq. (7), are summarized by the representation

$$\hat{\mu}(\sigma) = B_0(\sigma,E) + \hat{\mu}_1 B_1(\sigma,E) + \hat{\mu}_2 B_2(\sigma,E). \tag{13}$$

The $B_i(\sigma,E)$ coefficients are known and can be generated from Eq. (7) upon using the initial conditions $B_i(j) = \delta_{i,j}$, for $i,j = 0,1,2$.

Substitution of Eq. (13) into Eq. (12) results in

$$\tilde{\mu}(p) = \sum_{k=0}^{2} \hat{\mu}_k \Omega_k(p,E,C_0,...,C_I), \tag{14}$$

where $\hat{\mu}(0) \equiv 1$, and

$$\Omega_k(p,E,C_0,...,C_I) = \sum_{i,j=0}^{I} C_i^* C_j B_k(p+i+j,E). \tag{15}$$

We want to solve for the C's such that

$$\Omega_2(p,E,C_0,...,C_I) = 0, \quad \text{for } 0 \leqslant p \leqslant Q. \tag{16}$$

Clearly one expects the number of equations to be less than or equal to the number of variables, so

$$1 + Q \leqslant I. \tag{17}$$

If Eq. (16) can be solved, then the first $Q+1$ $\tilde{\mu}$ moments will depend on one less missing moment than the remaining $\tilde{\mu}$ moments, of order greater than Q.

Assuming the validity of Eq. (16), one can proceed to reduce the problem, inductively, to one of zero-missing moments, for the first $T+1$ moments. That is, assume a solution set $\{C\}$ to Eq. (16) has been determined. Let us define

$$\chi(x) \equiv \left|\sum_{l=0}^{L} D_l \cdot x^{2l}\right|^2 \phi(x). \tag{18}$$

The χ-Stieltjes moments satisfy

$$\hat{\mu}\chi(p) = \sum_{i,j=0}^{L} D_i^* D_j \tilde{\mu}(p+i+j). \tag{19}$$

Let T be a number satisfying $T + 2L \leqslant Q$. In accordance with Eqs. (14) and (16), we have

$$\hat{\mu}\chi(p) = \sum_{k=0}^{1} \hat{\mu}_k \Theta_k(p,E,D_0,...,D_L,C_0,...,C_I), \tag{20}$$

where $\hat{\mu}_0 \equiv 1$, $0 \leqslant p \leqslant T$, and

$$\Theta_k = \sum_{i,j=0}^{L} D_i^* D_j \Omega_k(p+i+j,E,C_0,...,C_I). \tag{21}$$

It is therefore clear that the solution of an equation analogous to that of Eq. (16), namely $\Theta_1 = 0$ (solving for the D's), will make the first $T+1$ χ-moments depend on no moments at all, besides being E dependent. Accordingly, we will focus on solving Eq. (16). The generalization of this inductive argument is immediate.

Equation (16) is equivalent to

$$\sum_{i,j=0}^{I} C_i^* C_j B_2(p+i+j,E) = 0, \quad 0 \leqslant p \leqslant Q. \tag{22}$$

The B's are known functions of E. From Padé analysis[4] it is always possible to find representations of the following type (E dependences are implicit):

$$B_2(\sigma) = \sum_{\nu=1}^{b} \alpha_\nu \beta_\nu^\sigma. \tag{23}$$

Equation (23) is an immediate consequence of the partial fraction decomposition (assuming a multiplicity of 1 for the roots) of an $[n/b]$ Padé approximant to the expansion

$$\sum_{\sigma=0}^{r} s^\sigma B_2(\sigma) = \frac{\sum_{\eta=0}^{n} s^\eta N(\eta)}{\sum_{\delta=0}^{b} s^\delta D(\delta)} + O(s^{1+r}) \tag{24}$$

$$= \sum_{\nu=1}^{b} \frac{\alpha_\nu}{1-\beta_\nu s}. \tag{25}$$

As usual, one requires $r = n + b$ and $n \leqslant b - 1$ [if Eq. (25) is to hold]. Note that in terms of I and Q, we have $r = Q + 2I$.

Insertion of Eq. (23) into Eq. (22) gives

$$\sum_{\nu=1}^{b} \alpha_\nu \beta_\nu^p \sum_{i,j=0}^{I} C_i^* C_j \beta_\nu^{i+j} = 0, \quad 0 \leqslant p \leqslant Q. \tag{26}$$

That is,

$$\sum_{\nu=1}^{b} \alpha_\nu \beta_\nu^p P(\beta_\nu) P^*(\beta_\nu^*) = 0, \quad 0 \leqslant p \leqslant Q, \tag{27}$$

where

$$P(\beta) = \sum_{i=0}^{I} C_i \beta^i. \tag{28}$$

It is emphasized that the β's are calculable functions of E! It is the C's that must be solved for!

The specific type of solution to Eq. (16), or Eq. (27), of interest are those in which the β's may be taken as roots of appropriate polynominals. From the Padé parameters it follows that if we set $n = b - \lambda$ $(1 \leqslant \lambda \leqslant b)$, then

$$b = I + (Q + \lambda)/2. \tag{29}$$

Because the degree of the P polynomial is I and there are b β's ($b > I$), clearly not all β's can be roots of this one polyno-

TABLE I. Generation of ground state energy bounds $E_- < E < E_+$, for quartic potential $v(x) = x^4$.

I	Q	λ	N_s	E_-	E_+
7	5	1	7	1.052	1.081
9	7	1	9	1.059 9	1.061 9
11	9	1	11	1.060 34	1.060 47
13	11	1	13	1.060 360	1.060 368
15	13	1	15	1.060 362 0	1.060 362 4
17	15	1	17	1.060 362 08	1.060 362 18

mial. However, a closer examination of Eq. (27) shows us that the roots of the P polynominal may correspond to

$$P(\beta_\nu) = 0, \quad \text{if } \beta_\nu \in \widetilde{S}, \tag{30}$$

where

$$\widetilde{S} \equiv \{\beta_\nu | \operatorname{Im} \beta_\nu \geqslant 0\}. \tag{31}$$

Accordingly, the C's are the polynomial coefficients for

$$\sum_{i=0}^{I} C_i s^i = \prod_{\beta_\nu \in \widetilde{S}} (s - \beta_\nu), \quad \text{if } N_s \leqslant I, \tag{32}$$

where N_s is the number of elements in $\widetilde{S}$.

The B_k's [refer to Eq. (23)] are not necessarily moments of a non-negative measure. Because of this, from the general theory of Stieltjes–Padé approximants it is to be expected that not all β's are real.[4] Accordingly, the number of elements in $\widetilde{S}$, N_s, should be small enough so that $N_s \leqslant I$. This is our basic assumption. Our expectation, confirmed by the quartic case to be presented, is that this is very likely to be true most of the time.

From the preceding discussion it is evident tha the realization of the basic inequality $N_s \leqslant I$, can best be achieved if b, the total number of β's, is as small as possible. Accordingly, it is best to start with $\lambda = b - n = 1$, the difference between the Padé approximant's denominator and numerator degrees.

IV. THE x^4 QUANTUM POTENTIAL

The quartic anharmonic oscillator $-\Psi'' + (mx^2 + x^4)\Psi = E\Psi$ is a one-missing moment problem with a recursive moment relation given by

$$\hat{\mu}(P+2) = [E\hat{\mu}(P) - m\hat{\mu}(P+1) + 2P(2P-1)\hat{\mu}(P-1)]. \tag{33}$$

Setting $\hat{\mu}(P) = B_0(P;E) + \hat{\mu}_1 B_1(P;E)$, we have

$$B_k(P+2) = [EB_k(P) - mB_k(P+1) + 2P(2P-1)B_k(P-1)], \tag{34}$$

where $B_i(j) = \delta_{i,j}$, for $i,j = 0,1$. Accordingly, the B_k's are computable functions of E. For the B_1's the following representation can be determined:

$$B_1(\sigma) = \sum_{\nu=1}^{b} \alpha_\nu \beta_\nu^\sigma, \quad 0 \leqslant \sigma \leqslant 2I + Q. \tag{35}$$

For simplicity, the remainder of this discussion will focus upon the massless case, $m = 0$. It will be noted that the $2P(2P-1)$ term in Eq. (34) may lead to very large B_k values, particularly if $2I + Q$ is large (of order 50). Because the HH inequalities are unaffected if we divide the $\hat{\mu}(P)$ moments by g^P, we may modify the B_k's accordingly, $\hat{B}_k(P) = B_k(P)/g^P$. Hence

$$\hat{B}_k(P+2) = (E/g^2)\hat{B}_k(P) + [2P(2P-1)/g^3]\hat{B}_k(P-1). \tag{36}$$

Clearly, $\beta_\nu \to \hat{\beta}_\nu \equiv \beta_\nu / g$. For the range of $2I + Q$ values quoted in Table I, an effective choice is $g = (2I + Q)/e$. This choice leads to moderate B_k values, which in turn lead to more accurate $\hat{\beta}$ values.

The program defined in the previous section was implemented. That is, the appropriate C's corresponding to Eq. (32) (for the β's) are determined. The first $Q + 1$ ϕ-Stieltjes moments [Eq. (11)] become dependent on E only:

$$\frac{\hat{\mu}(P)}{g^P} = \sum_{i,j=0}^{I} C_i(E)^* C_j(E) \hat{B}_0(P + i + j, E), \quad 0 \leqslant P \leqslant Q. \tag{37}$$

The application of the HH inequalities lead to constraints upon E. These are given in Table I. The results are consistent with the answer $E = 1.060\,362\,09$ from Ref. 1.

ACKNOWLEDGMENT

This work was supported in part by NSF Grant No. RII-8312.

[1] C. R. Handy and D. Bessis, Phys. Rev. Lett. **55**, 931 (1985).

[2] J. A. Shohat and J. D. Tamarkin, *The Problem of Moments* (Am. Math. Soc., Providence, RI, 1963).

[3] C. R. Handy, J. Phys. A **18**, 3593 (1985).

[4] G. Baker, Jr., *Essentials of Padé Approximants* (Academic, New York, 1975).

Factorization of the wave equation in a nonplanar stratified medium

Vaughan H. Weston
Department of Mathematics, Purdue University, West Lafayette, Indiana 47907

(Received 30 December 1986; accepted for publication 2 September 1987)

The wave equation is considered for a stratified medium where the stratifications are in the form of a family of nested C^2 surfaces along which the velocity c is constant (c varying only in a direction normal to the surfaces). On each surface c is constant, the solution u of the wave equation is decomposed into an outgoing wave component u^+ and an incoming wave component u^-. The associated outgoing and incoming wave conditions are expressed in terms of integral operators (kernels being time-dependent single and double layer potential type terms) operating on u and the normal derivative $\partial u/\partial n$ on each surface. Using the decomposition the scalar wave equation is split into a vector system involving the components u^+ and u^-, the vector system decoupling in a region where c is constant. Such a splitting is useful for the inverse problem where a reflection operator relating the outgoing wave to an incoming wave can be defined, and this in turn can be used to determine c.

I. INTRODUCTION

One of the techniques that has been used in the time-dependent direct and inverse scattering problems associated with the one-dimensional wave equation

$$\frac{\partial^2}{\partial z^2} u(z,t) = \frac{1}{c^2(z)} \frac{\partial^2}{\partial t^2} u(z,t), \quad z,t \in \mathbb{R}, \tag{1}$$

for a nonhomogeneous medium, is based upon the method of wave splitting.[1,2] By wave splitting we mean the decomposition of $u(z,t)$ into up-going (in the positive z direction) and down-going (in the negative z direction) waves. The importance of such splittings, in general, is that they lead to the use of invariant imbedding techniques.[3–6] Given a slab of inhomogeneous medium and a splitting one can define an associated scattering matrix. Invariant imbedding techniques then allow one to write a complex system of differential equations for the operator entries of the scattering matrix whose differentiation is with respect to the location of one of the planes of the slab. One can then deduce the behavior of the reflection operators for small time which provides a connection between up- and down-going wave fields and the properties of the medium on the edge of the slab.[1,2] The reflection operator can then be used in both direct and inverse scattering problems.

Various approaches[2,7–10] have been tried for extending the wave splitting to a planar stratified medium with $c = c(z)$ and $u = u(x,y,z,t)$. In particular, the approach taken by the author[10] was successful in giving rise to the form of the reflection operator and the explicit Ricatti type integral-differential equation and initial condition that the kernel of the reflection operator must satisfy.

The starting point of the procedure for wave splitting[10] in a planar stratified medium was the development of an upgoing and downgoing wave condition on a planar surface. This was obtained using Huygen's principle, or mathematically, by considering an initial value boundary-value problem for the wave equation in a homogeneous half-space. The resulting condition obtained for up-going and down-going waves on a surface $z = \text{const}$ is given by

$$u = \pm \mathscr{K}_0 u_z, \tag{2}$$

where

$$\mathscr{K}_0 v = -\int\!\!\int_{\mathbb{R}^2} \frac{v(x',y',t - R/c_0)}{2\pi R} H\left(t - \frac{R}{c_0}\right) dx'\, dy'. \tag{3}$$

(Here, the $+ve$ and $-ve$ signs refer to the up-going and down-going waves, respectively.)

This condition was then applied to decompose a solution $u(x,y,z,t)$ of the wave equation in a stratified medium into up- and down-going components by setting

$$u = u^+ + u^-, \quad u^\pm(x,y,z,t) = \frac{1}{2}\left(u \pm \mathscr{K} \frac{\partial u}{\partial z}\right), \tag{4}$$

where the operator $\mathscr{K}$ is the same as $\mathscr{K}_0$ with c_0 replaced by $c(z)$. The required system of equations satisfied by the components u^+ and u^- was obtained by first rewriting the wave equation in vector form

$$\frac{\partial}{\partial z}\begin{bmatrix} u \\ u_z \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ ((1/c^2)\,\partial_t^2 - \partial_x^2 - \partial_y^2) & 0 \end{bmatrix} \begin{bmatrix} u \\ u_z \end{bmatrix}, \tag{5}$$

and then transforming the system from one involving the dependent variables u, u_z to one involving the dependent variables u^+ and u^-. The resulting system took the form

$$\frac{\partial}{\partial z}\begin{bmatrix} u^+ \\ u^- \end{bmatrix} = W \begin{bmatrix} u^+ \\ u^- \end{bmatrix}, \tag{6}$$

where the matrix operator W is diagonal in a region where $c = \text{const}$. This system was subsequently used to obtain the equation for the kernel of the reflection operator $\mathscr{R}$, where $u^- = \mathscr{R} u^+$.

In this paper, it is shown that the approach for splitting developed previously[10] can be successfully extended for a nonplanar stratified medium in $\mathbb{R}^3$, where the decomposition is in terms of incoming and outgoing waves. In doing so, a number of key properties and identities of the time dependent single and double layer potentials which are of importance by themselves are established. The associated reflected operator that can be obtained from the resulting splitting is not investigated here.

A brief outline of the paper is as follows. In Sec. II, conditions for outgoing and incoming waves across the surface $\mathscr{S}$ in a homogeneous medium are obtained by considering the associated initial value exterior and interior problems for the homogeneous wave equation. The conditions involve the time-dependent single layer and double layer potentials and their normal derivatives. Operators $\mathfrak{K}$, $\mathfrak{M}$, $\mathfrak{N}$, $\mathfrak{L}$ are introduced that are associated with the potentials. In Sec. III, certain identities for these operators (which are needed in the splitting) are derived. The decomposition of a solution of the homogeneous wave equation into outgoing and incoming waves across $\mathscr{S}$ is then defined in Sec. IV. This definition is extended to the case where the medium is no longer homogeneous but is (nonplanar) stratified. In Sec. V, the existence of the inverse operator $\mathfrak{K}^{-1}$ is shown. This operator plays a key role and is necessary for the factorization of the wave equation in Sec. VI, where the generalization of the planar stratified splitting given by Eq. (6) is extended to a nonplanar stratified medium in $\mathbb{R}^3$. In Sec. VII, simplification of the splitting by dimensional reduction is given for two special nonplanar geometries. The details are presented for the more difficult case of the circular cylindrical geometry.

The following notation will be used in the remainder of the paper. Henceforth x and y will denote points in $\mathbb{R}^3$. Here G_i is a simply connected open region in $\mathbb{R}^3$ bounded by a (Lyapunov) C^2 surface $\mathscr{S}$. The surface $\mathscr{S}$ will either be a closed surface if the domain G_i is bounded or of infinite extent if the domain G_i is unbounded, but in either case it will have no edges. Examples of the former are spheres, ellipsoids, and of the latter are cylinders, planes. The surface $\mathscr{S}$ does not need to be convex. The associated exterior domain $\mathbb{R}^3\setminus\overline{G}_i$ will be denoted G_e. The unit normal on $\mathscr{S}$ directed outwards from G_i to G_e is given by n, with $\partial/\partial n_x$ and $\partial/\partial n_y$ being the corresponding normal derivatives at the points x and y on $\mathscr{S}$. When a function $f(x,y)$ of the two variables occurs, then the notation $\nabla_y f$, $\nabla_x f$ is used to denote the gradient with respect to y and x, respectively.

II. CONDITION FOR OUTGOING AND INCOMING WAVES ON A SURFACE $\mathscr{S}$ IMBEDDED IN A HOMOGENEOUS MEDIUM

Huygen's principle is employed to obtain the conditions for incoming and outgoing waves across a surface $\mathscr{S}$ in a medium with constant velocity c. These conditions [a generalization of Eq. (2), derived for a plane surface $\mathscr{S}$] will take on the form of a linear relationship between u and the normal derivative $\partial u/\partial n$ on $\mathscr{S}$. The outgoing wave condition will be obtained by considering the exterior initial value problem

$$\nabla^2 u - \frac{1}{c^2}\frac{\partial^2 u}{\partial t^2} = 0, \quad x\in G_e, \quad t>0, \tag{7a}$$

$$u = u_t = 0, \quad x\in\overline{G}_e, \quad t=0, \tag{7b}$$

where either u or $\partial u/\partial n$ are specified smooth C^2 functions on $\mathscr{S}$, which vanish for $t\leqslant 0$. System (7) can be placed in an integral formulation using the well-known Kirchhoff's formula,[11] valid for $x\in G_e$, $t\geqslant 0$,

$$u(x,t) = -\frac{1}{4\pi}\int_{\mathscr{S}}\left\{\frac{1}{R}u_n\left(y,t-\frac{R}{c}\right)+\frac{1}{R^2}\frac{\partial R}{\partial n_y}\right. \\ \left.\times\left[u\left(y,t-\frac{R}{c}\right)+\frac{R}{c}u_t\left(y,t-\frac{R}{c}\right)\right]\right\}d\sigma_y, \tag{8}$$

with $R=|x-y|$.

Then taking the limit as $x\in G_e\to x_0\in\mathscr{S}$, and employing the jump condition for the double-layer potential,[12,13] the following result is obtained:

$$(\mathfrak{I}+\mathfrak{M})u+\mathfrak{K}u_n=0, \quad x\in\mathscr{S}, \tag{9}$$

where the operators $\mathfrak{M}$ and $\mathfrak{K}$ are defined as follows:

$$\mathfrak{K}w[x,t\,] = \frac{1}{2\pi}\int_{\mathscr{S}}\frac{1}{R}u\left(y,t-\frac{R}{c}\right)H\left(t-\frac{R}{c}\right)d\sigma_y, \tag{10}$$

$$\mathfrak{M}w[x,t] = \frac{1}{2\pi}\int_{\mathscr{S}}\frac{1}{R^2}\frac{\partial R}{\partial n_y}\left[w\left(y,t-\frac{R}{c}\right)\right. \\ \left.+\frac{R}{c}w_t\left(y,t-\frac{R}{c}\right)\right]H\left(t-\frac{R}{c}\right)d\sigma_y. \tag{11}$$

Here $H(\eta)$ represents the Heaviside step function. The operators $\mathfrak{K}$ and $\mathfrak{M}$ are compact operators[14] with $\mathfrak{K}$ mapping $C(\mathscr{S})\times C[0,T]$ into itself, and $\mathfrak{M}$ mapping $C(\mathscr{S})\times C^1[0,T]$ into $C(\mathscr{S})\times C[0,T]$.

Relation (9) is the sought for outgoing wave condition on $\mathscr{S}$. For the case where $\mathscr{S}$ is a plane surface, the operator $\mathfrak{M}$ vanishes and this condition becomes identical to Eq. (2).

An alternative or reciprocal form of the outgoing condition (9) is obtained by taking the directional derivative $n_0\cdot\nabla_x$ of both sides of Eq. (8), where n_0 is the unit outward normal at the point x_0 on $\mathscr{S}$, and then taking the limit as $x\in G_e\to x_0\in\mathscr{S}$. Taking into account the jump in the normal derivative of a single layer potential, and the continuity of the normal derivative of a double layer potential with differentiable density[13,15] the resulting expression is given by

$$(\mathfrak{I}-\mathfrak{N})u_n+\mathfrak{L}u=0, \quad x\in\mathscr{S}. \tag{9'}$$

The compact[14] operator $\mathfrak{N}$ mapping $C(\mathscr{S})\times C^1[0,T]$ into $C(\mathscr{S})\times C[0,T]$ is defined by

$$\mathfrak{N}w[x,t] = \frac{1}{2\pi}\int_{\mathscr{S}}\frac{1}{R^2}\frac{\partial R}{\partial n_x}\left[w\left(y,t-\frac{R}{c}\right)\right. \\ \left.+\frac{R}{c}w_t\left(y,t-\frac{R}{c}\right)\right]H\left(t-\frac{R}{c}\right)d\sigma_y. \tag{12}$$

The operator $\mathfrak{L}$ is defined by

$$\mathfrak{L}w[x_0,t] = \lim_{x\in G_e\to x_0\in\mathscr{S}} n_0\cdot\nabla_x Q, \tag{13}$$

$$Q = \frac{1}{2\pi}\int_{\mathscr{S}}\frac{1}{R^2}\frac{\partial R}{\partial n_y} \\ \times\left[w\left(y,t-\frac{R}{c}\right)+\frac{R}{c}w_t\left(y,t-\frac{R}{c}\right)\right]H\left(t-\frac{R}{c}\right)d\sigma_y, \tag{13'}$$

where

$w(x,t)\in\{C^2(\mathscr{S})\times C^2[0,\infty)\cap w(y,0)=w_t(y,0)=0\}$.

An alternative form for $\mathfrak{L}$, expressed in terms of the tangential derivatives of a single-layer potential (which are continuous across $\mathscr{S}$) is given in Appendix A.

Equation (9′) constitutes a condition equivalent to (9) for outgoing waves on $\mathscr{S}$. Conditions (9) and (9′), developed for a medium of constant velocity c, will be modified later on to hold for a stratified medium where c varies in a direction normal to the surface $\mathscr{S}$.

The incoming wave conditions can be obtained in a similar manner by employing Kirchhoff's formula in region G_i. The resulting incoming wave conditions are given by

$$(\mathfrak{I}-\mathfrak{M})u-\mathfrak{K}u_n=0,\quad x\in\mathscr{S}, \tag{14}$$

$$(\mathfrak{I}+\mathfrak{N})u_n-\mathfrak{L}u=0,\quad x\in\mathscr{S}. \tag{14′}$$

Before employing these conditions to the factorization of the wave equation into incoming and outgoing waves, a number of identities involving the operators introduced in this section need to be established.

III. IDENTITIES INVOLVING THE OPERATORS $\mathfrak{K}, \mathfrak{M}, \mathfrak{N}, \mathfrak{L}$

A number of important identities among the operators can be easily obtained by considering two different representations of systems (7a) and (7b), one of these being the single layer potential type representation given by

$$u(x,t)=\frac{1}{2\pi}\int_{\mathscr{S}}\frac{1}{R}\,v\Big(y,t-\frac{R}{c}\Big)d\sigma_y, \tag{15}$$

and the other, the double layer potential type representation given by

$$u(x,t)=\frac{-1}{2\pi}\int_{\mathscr{S}}\frac{1}{R^2}\frac{\partial R}{\partial n_y}\times\Big[\mu\Big(y,t-\frac{R}{c}\Big)+\frac{R}{c}\mu_t\Big(y,t-\frac{R}{c}\Big)\Big]d\sigma_y. \tag{16}$$

All that is required of the densities $v(x,t)$ and $\mu(x,t)$ is that $v(x,t)=\mu(x,t)=0$ for $t\leqslant 0$, $\mu_t(x,0)=0$, and that

$$v(x,t)\in C(\mathscr{S})\times C^1[0,\infty)$$

and

$$\mu(x,t)\in C^2(\mathscr{S})\times C^2[0,\infty).$$

Taking the limit as $x\in G_e\to x\in\mathscr{S}$ in expressions (15) and (16) one obtains

$$u(x,t)=\mathfrak{K}v,\quad x\in\mathscr{S}, \tag{17}$$

$$u(x,t)=(\mathfrak{I}-\mathfrak{M})\mu,\quad x\in\mathscr{S}. \tag{18}$$

Taking the limit of the derivatives of expression (15) and (16) (in a manner indicated in the previous section), one obtains the resulting expressions for the normal derivative of u on $\mathscr{S}$,

$$u_n(x,t)=-(\mathfrak{I}+\mathfrak{N})v,\quad x\in\mathscr{S}, \tag{19}$$

$$u_n(x,t)=-\mathfrak{L}\mu,\quad x\in\mathscr{S}. \tag{20}$$

Now insert expressions (17) and (19) into outgoing wave conditions (9), to obtain

$$(\mathfrak{M}\mathfrak{K}-\mathfrak{K}\mathfrak{N})v=0. \tag{21}$$

Since $v(x,t)$ is an arbitrary function of $C(\mathscr{S})\times C^1[0,\infty)$ (vanishing at $t=0$), we immediately obtain

$$\mathfrak{M}\mathfrak{K}=\mathfrak{K}\mathfrak{N}. \tag{22}$$

This first identity is fairly obvious and can be easily verified by substituting in the appropriate expressions for the operators $\mathfrak{K}$, $\mathfrak{M}$, and $\mathfrak{N}$ into Eq. (21). In a similar manner, the insertion of expressions (18) and (20) into Eqs. (9) yields the following:

$$(\mathfrak{I}-\mathfrak{M}^2)\mu=\mathfrak{K}\mathfrak{L}\mu. \tag{23}$$

Since $\mu(x,t)$ is an arbitrary function of $C^2(\mathscr{S})\times C^2[0,\infty)$ we immediately obtain

$$\mathfrak{I}-\mathfrak{M}^2=\mathfrak{K}\mathfrak{L}. \tag{24}$$

Other identities can be obtained, such as the following:

$$(\mathfrak{I}-\mathfrak{N}^2-\mathfrak{L}\mathfrak{K})v=0, \tag{25}$$

however, this will require stronger conditions on the density $v(x,t)$. Since we will not be needing this particular result, we will not pursue it, other than to mention that when the surface $\mathscr{S}$ is a plane, the operators $\mathfrak{M}$ and $\mathfrak{N}$ vanish, and the resulting identities reduce to

$$\mathfrak{K}\mathfrak{L}=\mathfrak{L}\mathfrak{K}=\mathfrak{I},\quad \mathscr{S}\text{ a plane}, \tag{26}$$

a result obtained in the previous paper.[10]

In addition to the above identities, we need to consider the "normal derivative" of the operators $\mathfrak{K}$ and $\mathfrak{M}$ given by

$$\mathfrak{K}_n u=\frac{\partial}{\partial n}(\mathfrak{K}u), \tag{27}$$

$$\mathfrak{M}_n u=\frac{\partial}{\partial n}(\mathfrak{M}u), \tag{28}$$

with $u(y,t)$ as a function of y being defined on $\mathscr{S}$ only.

To obtain a precise form for the operators $\mathfrak{K}_n$, $\mathfrak{M}_n$, it will be assumed that $\mathscr{S}$ is a member of a nested family of surfaces and a curvilinear orthogonal coordinate system (ξ_1,ξ_2,ξ_3) can be chosen so that the surface $\mathscr{S}$ is given by $\xi_1=\xi_1^0$ (constant).

Let the points x and y on $\mathscr{S}$ have corresponding curvilinear coordinates (ξ_1,ξ_2,ξ_3) and (ξ_1',ξ_2',ξ_3') with $\xi_1=\xi_1'=\xi_1^0$. The element of area $d\sigma_y$ is given by

$$d\sigma_y=h_2'h_3'\,d\xi_2'\,d\xi_3', \tag{29}$$

where h_1', h_2', h_3' are the metric coefficients at the point of integration y. The metric coefficients at the point x will be denoted by h_1, h_2, h_3 (without primes).

A function $u(y,t)$ defined only on $\mathscr{S}\times[0,\infty)$ depends only on the transverse coordinate ξ_2', ξ_3' only and has the form $u(\xi_2',\xi_3',t)$.

Letting

$$f=(1/2\pi R)h_2'h_3'u(\xi_2',\xi_3',t-R/c),$$

the operator $\mathfrak{K}$ can be expressed in the general form

$$\mathfrak{K}u=\int\!\!\int f(x(\xi_1,\xi_2,\xi_3),y(\xi_1',\xi_2',\xi_3'))d\xi_2'\,d\xi_3',$$

where $\xi_1=\xi_1'=\xi_1^0$. With $\partial/\partial_n=(1/h_1)\partial/\partial\xi_1$, then

$$\mathfrak{K}_n u=\int\!\!\int\frac{1}{h_1}\frac{\partial f}{\partial\xi_1}d\xi_2'\,d\xi_3'=\int\!\!\int\Big[\frac{\partial f}{\partial n_x}+\frac{h_1'}{h_1}\frac{\partial f}{\partial n_y}\Big]d\xi_2'\,d\xi_3'. \tag{30}$$

Since

$$\frac{\partial R}{\partial n_x}+\frac{\partial R}{\partial n_y}=(n_x-n_y)\cdot\frac{(x-y)}{R},$$

it is seen that the indicated differentiation in expression (30) does not produce a higher order or worse singularity. It immediately follows on employing the relations

$$\frac{\partial f}{\partial n_x}=-\frac{1}{2\pi}\frac{1}{R^2}\frac{\partial R}{\partial n_x}\times\left[u\left(y,t-\frac{R}{c}\right)+\frac{R}{c}u_t\left(y,t-\frac{R}{c}\right)\right]h_2'h_3',$$

$$\frac{\partial f}{\partial n_y}=-\frac{1}{2\pi}\frac{1}{R^2}\frac{\partial R}{\partial n_y}\times\left[u\left(y,t-\frac{R}{c}\right)+\frac{R}{c}u_t\left(y,t-\frac{R}{c}\right)\right]h_2'h_3'+\frac{1}{2\pi R}u\left(y,t-\frac{R}{c}\right)\frac{\partial h_2'h_3'}{\partial n_y},$$

that

$$\mathfrak{N}_n u=-\mathfrak{N}u-\mathfrak{M}\left(\frac{h_1'}{h_1}u\right)+\frac{1}{h_1}\mathfrak{N}\left[u\frac{\partial}{\partial\xi_1'}\ln(h_2'h_3')\right].\tag{31}$$

Initially, the procedure for obtaining $\mathfrak{M}_n u$ is the same as for obtaining $\mathfrak{N}_n u$, however we require that $u(y,t)=u(\xi_2',\xi_3',t)$ be a twice differentiable function of the transverse variables ξ_2' and ξ_3'. As was done for $\mathfrak{N}u$, $\mathfrak{M}u$ can be placed in the form

$$\mathfrak{M}u=\int\int f(x(\xi_1,\xi_2,\xi_3),y(\xi_1',\xi_2',\xi_3'))d\xi_2'\,d\xi_3',$$

where $\xi_1'=\xi_1=\xi_1^0$ with

$$f=\frac{h_2'}{2\pi R^2}\frac{h_3'}{h_1'}\frac{\partial R}{\partial\xi_1'}\times\left[u\left(\xi_2',\xi_3',t-\frac{R}{c}\right)+\frac{R}{c}u_t\left(\xi_2',\xi_3',t-\frac{R}{c}\right)\right],$$

and then $\mathfrak{M}_n u$ is given by

$$\mathfrak{M}_n u=\int\int\left(\frac{\partial f}{\partial n_x}+\frac{h_1'}{h_1}\frac{\partial f}{\partial n_y}\right)d\xi_2'\,d\xi_3'.$$

Here the approach differs in that the integrals (with the point x on $\mathscr{S}$) are replaced by the corresponding integrals with $x\in G_e$ and followed by taking the limit as x approaches $\mathscr{S}$,

$$\mathfrak{M}_n u=\lim_{x\in G_e\to x\in\mathscr{S}}\int\int\left(\frac{\partial f}{\partial n_x}+\frac{h_1'}{h_1}\frac{\partial f}{\partial n_y}\right)d\xi_2'\,d\xi_3'.\tag{32}$$

The limits exist and is continuous (no jump discontinuity across $\mathscr{S}$ as will be shown). It is immediately seen that the limit of the first integral is $\mathfrak{L}u$, hence we have

$$\mathfrak{M}_n u-\mathfrak{L}u=\lim_{\xi_1\to\xi_1^0}\left(\frac{-J}{2\pi h_1}\right),\tag{33}$$

where

$$J=\int\int\frac{\partial}{\partial\xi_1'}\left\{\frac{h_2'h_3'}{h_1'}\left[u\left(\xi_2',\xi_3',t-\frac{R}{c}\right)+\frac{R}{c}u_t\left(\xi_2',\xi_3',t-\frac{R}{c}\right)\right]\frac{\partial 1/R}{\partial\xi_1'}\right\}d\xi_2'\,d\xi_3'.\tag{34}$$

The above integral is evaluated in Appendix B where it is shown to be

$$J=-\int\int_{\mathscr{S}}\frac{h_1'}{R}\times\left[\nabla_T^2u(y,\tau)-\frac{1}{c^2}\frac{\partial^2}{\partial\tau^2}u(y,\tau)\right]_{\tau=t-R/c}d\sigma_y,\tag{35}$$

where ∇_T^2 is the transverse Laplacian given by

$$\nabla_T^2=\frac{1}{h_1h_2h_3}\left\{\frac{\partial}{\partial\xi_2}\left(\frac{h_1h_3}{h_2}\frac{\partial}{\partial\xi_2}\right)+\frac{\partial}{\partial\xi_3}\left(\frac{h_1h_2}{h_3}\frac{\partial}{\partial\xi_3}\right)\right\}.\tag{36}$$

Combining terms we finally have

$$\mathfrak{M}_n u=\mathfrak{L}u+\frac{1}{h_1}\mathfrak{N}\left(h_1'\left[\nabla_T^2-\frac{1}{c^2}\frac{\partial^2}{\partial t^2}\right]u\right).\tag{37}$$

Using the above results, the following lemma is easily proved.

Lemma: Let D be an open domain in $\mathbf{R}^3$ containing $\mathscr{S}$, and $u(y,t)\in C^2(D)\times C^2[0,\infty)$, and let $u(y,t)=0$ for $t\leqslant 0$.

Then on $\mathscr{S}$,

$$\mathfrak{L}u-\mathfrak{N}\frac{\partial u}{\partial n}-\frac{\partial}{\partial n}\left(\mathfrak{M}u+\mathfrak{N}\frac{\partial u}{\partial n}\right)=-\mathfrak{N}\frac{h_1'}{h_1}\left(\nabla^2u-\frac{1}{c^2}\frac{\partial^2u}{\partial t^2}\right).\tag{38}$$

Proof: This immediately follows upon using the relations

$$\frac{\partial}{\partial n}\left(\mathfrak{M}u+\mathfrak{N}\frac{\partial u}{\partial n}\right)=\mathfrak{M}_nu+\left(\mathfrak{M}\frac{h_1'}{h_1}+\mathfrak{N}_n\right)\frac{\partial u}{\partial n}+\mathfrak{N}\left(\frac{h_1'}{h_1}\frac{\partial^2u}{\partial n^2}\right),$$

$$\nabla_T^2u+\frac{\partial^2u}{\partial n^2}+\frac{\partial\ln h_2h_3}{\partial n}\frac{\partial u}{\partial n}=\nabla^2u,$$

and the identities for $\mathfrak{M}_n$ and $\mathfrak{N}_n$ given by Eqs. (31) and (37).

IV. DECOMPOSITION OF WAVES IN A HOMOGENEOUS AND STRATIFIED MEDIUM

The decomposition of waves into outgoing and incoming waves across a closed surface $\mathscr{S}$ imbedded in a homogeneous medium can now be defined

$$u^+=\frac{1}{2}\left[u-\mathfrak{M}u-\mathfrak{N}\frac{\partial u}{\partial n}\right],\tag{39}$$

$$u^-=\frac{1}{2}\left[u+\mathfrak{M}u+\mathfrak{N}\frac{\partial u}{\partial n}\right].\tag{40}$$

Lemma: If u satisfies the wave equation in a homogeneous medium characterized by an open domain $D\in\mathbf{R}^3$ containing the surface $\mathscr{S}$, then u can be decomposed into outgoing waves u^+ and incoming waves u^-, by the relation

$$u=u^++u^-.\tag{41}$$

Proof: Since the decomposition given by Eq. (41) is obvious, the critical thing is to establish that u^+ is an outgoing wave across $\mathscr{S}$. (The corresponding result for u^- is then similarly established.)

Since

$$\begin{aligned}
&2(\mathfrak{I}+\mathfrak{M})u^+ + 2\mathfrak{R}\frac{\partial u^+}{\partial n}\\
&\quad=(\mathfrak{I}-\mathfrak{M}^2)u-\mathfrak{M}\mathfrak{R}\frac{\partial u}{\partial n}\\
&\qquad-\mathfrak{R}\frac{\partial}{\partial n}(\mathfrak{M}u)-\mathfrak{R}\frac{\partial}{\partial n}\left(\mathfrak{R}\frac{\partial u}{\partial n}\right)\\
&\quad=\mathfrak{R}\left[\mathfrak{L}u-\mathfrak{N}\frac{\partial u}{\partial n}-\frac{\partial}{\partial n}(\mathfrak{M}u)-\frac{\partial}{\partial n}\left(\mathfrak{R}\frac{\partial u}{\partial n}\right)\right],
\end{aligned}$$

using relations (22) and (24), it follows from identity (38) that because u satisfies the wave equation in a homogeneous medium, that

$$(\mathfrak{I}+\mathfrak{M})u^+ + \mathfrak{R}\frac{\partial u^+}{\partial n}=0,$$

which is the outgoing wave condition (9), thus proving the required result. ■

The concept of the decomposition into outgoing and incoming waves will be extended to a stratified medium, the stratification being described by a set of nested closed nonoverlapping surfaces, with the velocity c being constant along each surface. Employing a curvilinear orthogonal coordinate system (ξ_1,ξ_2,ξ_3) the stratification will be explicitly specified by setting $c=c(\xi_1)$.

The extension of the incoming and outgoing wave concept to a stratified medium will be done in a manner similar to previous work,[10] namely by first thinking of the stratification as a set of thin homogeneous layers of finite thickness. Then the outgoing and incoming wave conditions (9) and (14) and decomposition (41) is applied in each homogeneous layer. Finally, the layers are allowed to have infinitesimal thickness. What this implies is that the operators $\mathfrak{R}$, $\mathfrak{M}$, $\mathfrak{N}$, $\mathfrak{L}$ will be modified for a stratified medium by requiring the surface $\mathscr{S}$ specified by coordinate $\xi_1=\xi_1^0$ (constant), in which case $c=c(\xi_1)$ (c remains constant along $\mathscr{S}$). For example, the operator $\mathfrak{R}$ will be given by

$$\mathfrak{R}u=\frac{1}{2\pi}\int_{\mathscr{S}}\frac{1}{R}u\left(y,t-\frac{R}{c(\xi_1)}\right)d\sigma_y,\quad x\in\mathscr{S}(\xi_1=\xi_1^0).$$

Since c is still constant along $\mathscr{S}$, the identities (22) and (24), applied to surface integrals over $\mathscr{S}$, will still hold. The only change that will occur is in the normal derivative of the operators $\mathfrak{R}$ and $\mathfrak{M}$. With $u(y,t)$ being a C^2 function in an open region containing $\mathscr{S}$ and which vanishes for $t\leqslant 0$, the normal derivative of $\mathfrak{R}u$ is easily obtained to be

$$\frac{\partial}{\partial n}(\mathfrak{R}u)=\mathfrak{R}_n u+\frac{1}{h_1}\mathfrak{R}\left(h_1'\frac{\partial u}{\partial n}\right)+\frac{\partial c}{\partial n}\mathfrak{R}_c u,\qquad(42)$$

with

$$\begin{aligned}
\mathfrak{R}_c u&=\frac{1}{2\pi c^2}\int_{\mathscr{S}}u_t\left(y,t-\frac{R}{c}\right)H\left(t-\frac{R}{c}\right)d\sigma_y\\
&=\frac{t}{c}\mathfrak{R}u_t-\frac{1}{c}\mathfrak{R}(tu_t),\qquad(42')
\end{aligned}$$

and $\mathfrak{R}_n$ given by Eq. (31). In a similar manner it can be shown that

$$\frac{\partial(\mathfrak{M}u)}{\partial n}=\mathfrak{M}_n u+\frac{1}{h_1}\mathfrak{M}\left(h_1'\frac{\partial u}{\partial n}\right)+\frac{\partial c}{\partial n}\mathfrak{M}_c u,\qquad(43)$$

with

$$\begin{aligned}
\mathfrak{M}_c u&=\frac{1}{2\pi c^3}\int_{\mathscr{S}}\frac{\partial R}{\partial n_y}u_{tt}\left(y,t-\frac{R}{c}\right)H\left(t-\frac{R}{c}\right)d\sigma_y\\
&=\frac{t}{c}\mathfrak{M}u_t-\frac{1}{c}\mathfrak{M}(tu_t),\qquad(43')
\end{aligned}$$

and $\mathfrak{M}_n$ given by Eq. (37).

V. INVERSE OPERATOR $\mathfrak{R}^{-1}$

It was shown in the previous paper[10] that the inverse operator $\mathfrak{R}^{-1}$ exists when $\mathscr{S}$ is a plane surface, and its explicit form was derived. Here we examine the question of existence of $\mathfrak{R}^{-1}$ for the general case where $\mathscr{S}$ is a smooth closed surface.

Since $\mathfrak{R}$ is a compact operator[14] mapping $C(\mathscr{S})\times C[0,T]$ into $C(\mathscr{S})\times C[0,T]$, all that is needed to be shown is that the null space of $\mathfrak{R}$ is empty. It immediately follows then $\mathfrak{R}^{-1}$ exists.

A brief outline of the proof that the null space of $\mathfrak{R}$ is empty, is as follows (for further details on this and some more general results see Ref. 14).

Let $v(x,t)\in C(\mathscr{S})\times C[0,T]$ be a solution of $\mathfrak{R}v=0$. Then set

$$u(x,t)=\frac{1}{2\pi}\int_{\mathscr{S}}\frac{1}{R}v\left(y,t-\frac{R}{c}\right)H\left(t-\frac{R}{c}\right)d\sigma_y,\quad x\in\mathbb{R}^3.$$

It immediately follows that $u(x,t)$ is a solution of the mixed problem in the exterior domain G_e, satisfying the wave equation and the initial conditions $u(x,0)=u_t(x,0)=0$, $x\in G_e$, as well as the Dirichlet boundary condition $u=0$ on $\mathscr{S}$. It follows[11] that the solution $u(x,t)\equiv 0$, for $x\in G_e$, $t>0$. A similar result can be deduced from the interior domain G_i. It thus follows from the jump condition $[\partial u/\partial n]_-^+=-2v(x,t)$, $x\in\mathscr{S}$, that $v\equiv 0$. The result is summarized as follows.

Lemma: The null space of $\mathfrak{R}$ is empty, and $\mathfrak{R}^{-1}$ exists.

VI. FACTORIZATION OF THE WAVE EQUATION IN A STRATIFIED MEDIUM

Here we will consider the factorization of the wave equation

$$\begin{aligned}
&\frac{1}{c^2(\xi_1)}\frac{\partial^2 u}{\partial t^2}=\nabla^2 u,\quad x\in D,\quad t>0,\\
&u(x,0)=0,\quad x\in D,\quad t\leqslant 0,
\end{aligned}\qquad(44)$$

for a stratified medium [stratification described by surfaces $\xi_1=\xi_1(x,y,z)=\text{const}$]. We will restrict ourselves to a region D so that the requirement $u\equiv 0$, for $t\leqslant 0$ can be imposed here. Thus any sources producing the wave will lie outside D. In practical applications, D would be a scattering body or a portion of it.

The decomposition into incoming and outgoing waves given by Eqs. (39) and (40) will be expressed in vector form as follows:

$$\begin{bmatrix} u^+ \\ u^- \end{bmatrix} = T \begin{bmatrix} u \\ \partial u/\partial n \end{bmatrix}, \tag{45}$$

where the matrix operator T is given by

$$T = \frac{1}{2} \begin{bmatrix} (\mathfrak{J} - \mathfrak{M}) & -\mathfrak{R} \\ (\mathfrak{J} + \mathfrak{M}) & \mathfrak{R} \end{bmatrix}. \tag{46}$$

The inverse of T is given by

$$T^{-1} = \begin{bmatrix} 1 & 1 \\ -\mathfrak{R}^{-1}(\mathfrak{J} + \mathfrak{M}) & \mathfrak{R}^{-1}(\mathfrak{J} - \mathfrak{M}) \end{bmatrix}. \tag{47}$$

The wave equation expressed in curvilinear coordinate system (ξ_1, ξ_2, ξ_3) will be written in the form

$$\frac{1}{h_1 h_2 h_3} \frac{\partial}{\partial \xi_1} \left(\frac{h_2 h_3}{h_1} \frac{\partial u}{\partial \xi_1} \right) = \Box_T u, \tag{48}$$

where

$$\Box_T = \frac{1}{c^2(\xi_1)} \frac{\partial^2}{\partial t^2} - \nabla_T^2, \tag{49}$$

with ∇_T^2 being the transverse component of the Laplacian given by Eq. (36). The combination of Eq. (48) with the trivial identity

$$\frac{1}{h_1 h_2 h_3} \frac{\partial u}{\partial \xi_1} = \frac{1}{h_1 h_2 h_3} \frac{\partial u}{\partial \xi_1}$$

results in the following system expressed in vector form:

$$\frac{1}{h_1 h_2 h_3} \frac{\partial}{\partial \xi_1} \begin{bmatrix} u \\ (h_2 h_3/h_1)(\partial u/\partial \xi_1) \end{bmatrix} = \begin{bmatrix} 0 & (h_2 h_3)^{-2} \\ \Box_T & 0 \end{bmatrix} \begin{bmatrix} u \\ (h_2 h_3/h_1)(\partial u/\partial \xi_1) \end{bmatrix}. \tag{50}$$

System (50) will be expressed in terms of incoming and outgoing waves (changing the basis), by first inverting system (45) to yield

$$\begin{bmatrix} u \\ (1/h_1)(\partial u/\partial \xi_1) \end{bmatrix} = T^{-1} \begin{bmatrix} u^+ \\ u^- \end{bmatrix}, \tag{51}$$

then inserting expression (51) into Eq. (50), performing the necessary differentiation, and then premultiplying the resulting system by the matrix

$$T \begin{bmatrix} h_1 h_2 h_3 & 0 \\ 0 & h_1 \end{bmatrix} \tag{52}$$

to obtain

$$\frac{\partial}{\partial \xi_1} \begin{bmatrix} u^+ \\ u^- \end{bmatrix} = W \begin{bmatrix} u^+ \\ u^- \end{bmatrix}, \tag{53}$$

with

$$W = T \begin{bmatrix} 0 & h_1 \\ h_1 \Box_T & -\partial/\partial \xi_1 \ln(h_2 h_3) \end{bmatrix} T^{-1} - T \frac{\partial T^{-1}}{\partial \xi_1}. \tag{54}$$

Since it is easier to differentiate T than T^{-1}, the second term of W will be rewritten using the relation

$$0 = \frac{\partial}{\partial \xi_1} (T T^{-1}) = T \left(\frac{\partial T^{-1}}{\partial \xi_1} \right) + \left(\frac{\partial T}{\partial \xi_1} \right) T^{-1}$$

together with the following expression for the components of the operator $\partial T/\partial \xi_1$:

$$\left(\frac{\partial \mathfrak{R}}{\partial \xi_1} \right) u = h_1 \left\{ \mathfrak{R}_n u + \frac{\partial c}{\partial n} \mathfrak{R}_c u \right\},$$

$$\left(\frac{\partial \mathfrak{M}}{\partial \xi_1} \right) u = h_1 \left\{ \mathfrak{M}_n u + \frac{\partial c}{\partial n} \mathfrak{M}_c u \right\},$$

obtained from Eqs. (42) and (43) with u independent of the variable ξ_1. This yields

$$-T \left(\frac{\partial T^{-1}}{\partial \xi_1} \right) = \frac{1}{2} h_1 \begin{bmatrix} -\left(\mathfrak{M}_n + \frac{\partial c}{\partial n} \mathfrak{M}_c \right) & -\left(\mathfrak{R}_n + \frac{\partial c}{\partial n} \mathfrak{R}_c \right) \\ \left(\mathfrak{M}_n + \frac{\partial c}{\partial n} \mathfrak{M}_c \right) & \left(\mathfrak{R}_n + \frac{\partial c}{\partial n} \mathfrak{R}_c \right) \end{bmatrix} T^{-1}. \tag{55}$$

With the insertion of expression (55) into relation (54) and employment of expressions (31) and (37) for the operators $\mathfrak{R}_n$ and $\mathfrak{M}_n$, respectively, the following simplified expression for W is obtained:

$$W = \frac{h_1}{2} \begin{bmatrix} -\mathfrak{L} & (\mathfrak{J} + \mathfrak{N}) \\ \mathfrak{L} & (\mathfrak{J} - \mathfrak{N}) \end{bmatrix} T^{-1} + \frac{1}{2} h_1 \frac{\partial c}{\partial n} \begin{bmatrix} -\mathfrak{M}_c & -\mathfrak{R}_c \\ \mathfrak{M}_c & \mathfrak{R}_c \end{bmatrix} T^{-1}. \tag{56}$$

By using the identities [obtained from relations (22) and (24)]

$$(\mathfrak{J} \pm \mathfrak{N}) \mathfrak{R}^{-1} (\mathfrak{J} \pm \mathfrak{M}) + \mathfrak{L} = \mathfrak{R}^{-1} [(\mathfrak{J} \pm \mathfrak{M})^2 + \mathfrak{R} \mathfrak{L}] = 2 \mathfrak{R}^{-1} (\mathfrak{J} \pm \mathfrak{M}), \tag{57}$$

$$(\mathfrak{J} \pm \mathfrak{N}) \mathfrak{R}^{-1} (\mathfrak{J} \mp \mathfrak{M}) - \mathfrak{L} = \mathfrak{R}^{-1} [\mathfrak{J} - \mathfrak{M}^2 - \mathfrak{R} \mathfrak{L}] = 0, \tag{58}$$

the first term in expression (56) can be multiplied out and simplified to give

$$W = h_1 \begin{bmatrix} -\mathfrak{R}^{-1}(\mathfrak{J} + \mathfrak{M}) & 0 \\ 0 & \mathfrak{R}^{-1}(\mathfrak{J} - \mathfrak{M}) \end{bmatrix} + \frac{1}{2} \frac{\partial c}{\partial \xi_1} \begin{bmatrix} -\mathfrak{M}_c & -\mathfrak{R}_c \\ \mathfrak{M}_c & \mathfrak{R}_c \end{bmatrix} T^{-1}. \tag{59}$$

It is immediately apparent that when c is a constant, w becomes a diagonal matrix, and the outgoing and incoming waves are decoupled.

The results are collected on the following.

Theorem: For a stratified medium where the velocity c is a function of the coordinate ξ_1, the solution u of the system (44) can be split up into incoming waves u^- [defined by Eq. (40)] and outgoing waves u^+ [defined by Eq. (39)] and these components are related through the system

$$\frac{\partial}{\partial \xi_1} \begin{bmatrix} u^+ \\ u^- \end{bmatrix} = h_1 \begin{bmatrix} -\mathfrak{R}^{-1}(\mathfrak{J} + \mathfrak{M}) & 0 \\ 0 & \mathfrak{R}^{-1}(\mathfrak{J} - \mathfrak{M}) \end{bmatrix} \begin{bmatrix} u^+ \\ u^- \end{bmatrix} + \frac{1}{2} \frac{\partial c}{\partial \xi_1} \begin{bmatrix} -\mathfrak{M}_c & -\mathfrak{R}_c \\ \mathfrak{M}_c & \mathfrak{R}_c \end{bmatrix} T^{-1} \begin{bmatrix} u^+ \\ u^- \end{bmatrix}. \tag{60}$$

System (60) is the sought for result in this paper. It clearly demonstrates that the wave splitting concept can be extended to nonplanar stratified medium.

To complete the analysis, the form of the reflection operator $\mathscr{R}$ that relates the outgoing wave to the incoming wave u^- by the relation $u^+ = \mathscr{R} u^-$, and the equation satisfied by the kernel of the integral operator $\mathscr{R}$, need to be established. For the one-dimensional problem, the form of the reflection operator is given in terms of a simple convolution. However, for the planar stratified case developed in the previous paper,[10] it took on a more complex form involving an operator and a convolution. This indicates that it would be easier to ascertain the form of the reflection operator for some special cases, and then generalize the results to a general stratified medium. Thus before investigating the general case, it would be more prudent to examine system (60) for the cases of spherical or circular cylindrical geometry. Inverse problems involving spherical or cylindrical stratified medium can be reduced to one spatial-dimensional problem when the scattered field is measured over a spherical or cylindrical surface. Because this restricted class of problems is of interest the associated reduced form of the wave splitting will be presented here. This is given in the next section with emphasis on the more difficult case of the circular cylindrical geometry.

VII. REDUCTION TO ONE-SPATIAL DIMENSION

For the case where the stratified surfaces have the property that $(\partial/\partial\xi_1)\ln(h_2h_3)$ and h_1 are independent of ξ_2 and ξ_3, the multidimensional problem is reducible to a one-dimensional spatial problem. The stated conditions imply that the surface $\mathscr{S}(\xi_1=\text{const})$ has constant mean curvature and that the coordinate curves orthogonal to it are straight. Two such systems that have this property are (i) the spherical polar coordinate system (ρ,θ,ϕ), and (ii) the cylindrical polar coordinate system (ρ,θ,x_3), where in both cases $\xi_1=\rho$.

By noting that the above conditions have the explicit form

$$h_2h_3=f(\xi_1)g(\xi_2\xi_3),\quad h_1=h_1(\xi_1), \tag{61}$$

the reduction to the one-dimensional problem is obtained by multiplying Eq. (44) by $1/f(\xi_1)$ and integrating over the surface $\mathscr{S}$, yielding

$$\frac{1}{h_1f}\frac{\partial}{\partial\xi_1}\left(\frac{f}{h_1}\frac{\partial v}{\partial\xi_1}\right)=\frac{1}{c^2(\xi_1)}\frac{\partial^2 v}{\partial t^2}, \tag{62}$$

where

$$v=\mathfrak{S}u, \tag{63}$$

with the operator $\mathfrak{S}$ defined by

$$\mathfrak{S}u=\int_{\mathscr{S}}u(\xi_1,\xi_2,\xi_3,t)g(\xi_2,\xi_3)d\xi_2\,d\xi_3. \tag{64}$$

For the spherical polar case (where $\xi_1=\rho$, $h_1=1$, $f=\rho^2$), the factorization of Eq. (62) can be obtained directly by setting $v(\rho,t)=w(\rho,t)/\rho$ thus reducing Eq. (62) to a one-dimensional wave equation involving $w(\rho,t)$ for which the factorization is well established.[1,2] Of more interest then is the case of cylindrical polar coordinates (where $\xi_1=\rho$, $h_1=1,f=\rho$). The corresponding factorization for cylindrical polar coordinates (ρ,θ,x_3) is obtained from Eq. (60) by operating on both components of this system with the operator $\mathfrak{S}$. To get the appropriate form for the factorization, we need to examine the expressions of $\mathfrak{S}\mathfrak{R}u$, $\mathfrak{S}\mathfrak{M}u$, $\mathfrak{S}\mathfrak{R}^{-1}u$, etc.

Rewriting expression for the operator $\mathfrak{R}$ in the form

$$\mathfrak{R}u[x,t]=c\int_{\mathscr{S}}\int_0^\infty\frac{\delta(c(t-s)-R)}{2\pi R}u(y,s)d\sigma_y\,ds,$$

where $R=|x-y|$ with x and y being points on the surface $\mathscr{S}$ ($\rho=\text{const}$) and $d\sigma_y=\rho\,d\theta_y\,dy_3$, it is seen that with $g\equiv1$ in expression (64) that

$$\mathfrak{S}\mathfrak{R}u=\int_{\mathscr{S}}\int_0^\infty\left\{\int_{\mathscr{S}}\frac{c\delta(c(t-s)-R)}{2\pi R}d\sigma_x\right\}\frac{u(y,s)}{\rho}d\sigma_y\,ds.$$

It is shown in Appendix C that

$$\int_{\mathscr{S}}c\frac{\delta(ct-R)}{2\pi R}d\sigma_x=k(\rho,t)H(t), \tag{65}$$

where

$$k(\rho,t)=\begin{cases}(2c/\pi)K(\zeta), & 0\leqslant t<2\rho/c,\\ (4\rho/\pi t)K(1/\zeta), & 2\rho/c\leqslant t,\end{cases} \tag{66}$$

with

$$\zeta=(ct/2\rho), \tag{67}$$

and $K(\zeta)$ is the complete (Legendre) elliptic integral of the first kind[16]:

$$K(\zeta)=\int_0^1\frac{1}{\sqrt{(1-\eta^2)(1-\zeta^2\eta^2)}}d\eta. \tag{68}$$

It follows then that

$$\mathfrak{S}\mathfrak{R}u=\hat{\mathfrak{R}}\mathfrak{S}u, \tag{69}$$

where

$$\hat{\mathfrak{R}}v=\int_0^t k(\rho,t-s)v(\rho,s)ds. \tag{70}$$

Note that the kernel $k(\rho,t)$ has a logarithmic singularity when $ct=2\rho$.

The inverse of the operator $\hat{\mathfrak{R}}$ can be easily obtained for the time $0\leqslant t<2\rho/c$ by differentiating the equation

$$\hat{\mathfrak{R}}v=w,$$

with t to give

$$v(\rho,t)+\int_0^t\frac{1}{c}k_t(\rho,t-s)v(\rho,s)ds=\frac{1}{c}\frac{\partial w}{\partial t}.$$

Since this constitutes a Volterra integral equation of the second kind with continuous kernel $(0\leqslant t<2\rho/c)$, it can be solved by iterations to yield

$$v=(I+\hat{\mathfrak{H}})^{-1}\frac{1}{c}\frac{\partial w}{\partial t}$$
$$=\sum_{n=0}^{\infty}(-1)^n(\hat{\mathfrak{H}})^n\frac{1}{c}\frac{\partial w}{\partial t},\quad 0\leqslant t<\frac{2\rho}{c},$$

where

$$\hat{\mathfrak{H}}v=\int_0^t\frac{1}{c}k_t(\rho,t-s)v(\rho,s)ds. \tag{71}$$

Thus it follows that

$$\hat{\mathfrak{R}}^{-1}=(I+\hat{\mathfrak{H}})^{-1}\frac{1}{c}\frac{\partial}{\partial t},\quad 0\leqslant t<\frac{2\rho}{c}. \tag{72}$$

The form of the operator $\mathfrak{S}\,\mathfrak{R}^{-1}$ can now be obtained by operating on both sides of the identity $\mathfrak{R}\,\mathfrak{R}^{-1}u=u$ with $\mathfrak{S}$, employing relation (69) to give $\hat{\mathfrak{R}}\,\mathfrak{S}\,\mathfrak{R}^{-1}u=\mathfrak{S}u$, and finally inverting to obtain $\mathfrak{S}\,\mathfrak{R}^{-1}u=\hat{\mathfrak{R}}^{-1}\mathfrak{S}u$. This yields the relation

$$\mathfrak{S}\mathfrak{R}^{-1}=\hat{\mathfrak{R}}^{-1}\mathfrak{S}. \tag{73}$$

The evaluation of $\mathfrak{S}\,\mathfrak{M}$ follows the same way as for $\mathfrak{S}\,\mathfrak{R}$. From Eq. (11), $\mathfrak{S}\mathfrak{M}u$ can be written in the form

$$\mathfrak{S}\mathfrak{M}u=\int_{\mathscr{S}}\int_0^\infty\Big\{m_2(\rho,t-s)u(y,s)+\frac{1}{c}m_1(\rho,t-s)u_s(y,s)\Big\}\frac{1}{\rho}\,ds\,d\sigma_y, \tag{74}$$

where

$$m_j(\rho,t)=\frac{c}{2\pi}\int_{\mathscr{S}}\frac{1}{R^j}\frac{\partial R}{\partial n_y}\delta(ct-R)d\sigma_x,\quad j=1,2, \tag{75}$$

which are evaluated in Appendix C [see Eqs. (C3) and (C4)]. Using these results Eq. (74) can be expressed in the following:

$$\mathfrak{S}\mathfrak{M}u=\int_0^t m_2(\rho,t-s)\,[v(\rho,s)+(t-s)v_s(\rho,s)\,]ds, \tag{76}$$

where

$$v=\mathfrak{S}u. \tag{77}$$

For the class of functions $u(y,t)$ such that $u(y,0)=0$, expression (76) can be integrated by parts to give

$$\mathfrak{S}\mathfrak{M}u=\hat{\mathfrak{M}}v=\int_0^t m(\rho,t-s)v(\rho,s)ds, \tag{78}$$

where

$$m(\rho,t)=2m_2(\rho,t)+t\frac{\partial}{\partial t}m_2(\rho,t).$$

This is evaluated in Appendix C as

$$m(\rho,t)=(c/\pi\rho)[K(\zeta)+\zeta K'(\zeta)],\quad 0\leqslant t<2\rho/c, \tag{79}$$

with ζ given by Eq. (67).

Finally, the operators $\mathfrak{S}\mathfrak{R}_c$ and $\mathfrak{S}\mathfrak{M}_c$ can be obtained. From Eqs. (42′), (69), (70) it follows that

$$\mathfrak{S}\mathfrak{R}_c u=\int_0^t\frac{(t-s)}{c}k(\rho,t-s)v_s(\rho,s)ds,$$

which, on integrating by parts, yields for functions $u(y,t)$ that vanish at $t=0$,

$$\mathfrak{S}\mathfrak{R}_c u=(2\rho/c)\hat{\mathfrak{M}}v,\quad 0\leqslant t<2\rho/c,$$

where v is related to u by Eq. (77). This implies then that

$$\mathfrak{S}\mathfrak{R}_c=(2\rho/c)\hat{\mathfrak{M}}\mathfrak{S}. \tag{80}$$

The following additional result follows from Eqs. (43′) and (78):

$$\mathfrak{S}\mathfrak{M}_c=\hat{\mathfrak{M}}_c\mathfrak{S}, \tag{81}$$

where

$$\hat{\mathfrak{M}}_c v=\frac{1}{c}\int_0^t\{m(\rho,t-s)+(t-s)m_t(\rho,t-s)\}v(\rho,s)ds, \tag{82}$$

for $0\leqslant t<2\rho/c$.

The factorization of the reduced form of the wave equation [Eq. (62)] in cylindrical polar coordinates can now be obtained directly by operating on components of system (60) with $\mathfrak{S}$. Setting

$$v^+=\mathfrak{S}u^+,\quad v^+=\mathfrak{S}u^-, \tag{83}$$

the resulting factorized equations are

$$\frac{\partial}{\partial\rho}\begin{bmatrix}v^+\\ v^-\end{bmatrix}=\begin{bmatrix}-\hat{\mathfrak{R}}^{-1}(\mathfrak{J}+\hat{\mathfrak{M}}) & 0\\ 0 & \hat{\mathfrak{R}}^{-1}(\mathfrak{J}-\hat{\mathfrak{M}})\end{bmatrix}\begin{bmatrix}v^+\\ v^-\end{bmatrix}+\frac{1}{2}\frac{\partial c}{\partial\rho}\begin{bmatrix}\mathfrak{A} & \mathfrak{B}\\ -\mathfrak{A} & -\mathfrak{B}\end{bmatrix}\begin{bmatrix}v^+\\ v^-\end{bmatrix}, \tag{84}$$

where

$$\mathfrak{A}=-\hat{\mathfrak{M}}_c+(2\rho/c)\hat{\mathfrak{M}}\hat{\mathfrak{R}}^{-1}(\mathfrak{J}+\hat{\mathfrak{M}}),$$
$$\mathfrak{B}=-\hat{\mathfrak{M}}_c-(2\rho/c)\hat{\mathfrak{M}}\hat{\mathfrak{R}}^{-1}(\mathfrak{J}-\hat{\mathfrak{M}}).$$

The splitting given by Eq. (84) is extremely useful for inverse problems involving cylindrical geometry (even when the incident wave is produced by a point source). What is needed is the form of the reflection operator relating the outgoing wave to the incoming wave, and the equation for the kernel of the reflection operator. This will be done in a subsequent paper where in addition the reflection operator will be employed to solve a class of inverse problems.

ACKNOWLEDGMENT

The research was supported by O.N.R. under Contract No. N 00014-83-K-6038.

APPENDIX A: ALTERNATIVE EXPRESSION FOR $\mathfrak{L}\boldsymbol{w}[\boldsymbol{x},\boldsymbol{t}]$

To obtain an alternative form for $\mathfrak{L}$, the integral Q [given by Eq. (13′)] will be written in the form

$$Q=\int_{\mathscr{S}}H(t^*)(n_y\cdot\nabla_x)\left[\frac{w(y,t^*)}{2\pi R}\right]d\sigma_y, \tag{A1}$$

where $t^*=t-R/c$. (A2)

With $w(x,t)$ a twice differentiable function such that $w(y,0)=w_t(y,0)=0$, it can be shown that for $x\in G_e$,

$$\nabla_x Q=\int_{\mathscr{S}}H(t^*)(n_y\cdot\nabla_x)\nabla_x\left[\frac{w(y,t^*)}{2\pi R}\right]d\sigma_y. \tag{A3}$$

This is reduced using the identity

$$(n_y\cdot\nabla_x)\nabla_x\left[\frac{w(y,t^*)}{R}\right]=n_y\nabla_x^2\left[\frac{w(y,t^*)}{R}\right]-\nabla_x\times\left(n_y\times\nabla_x\left[\frac{w(y,t^*)}{R}\right]\right), \tag{A4}$$

and the fact that the first term on the right-hand side of Eq. (A4) is $n_y w_{tt}(y,t^*)/(c^2R)$ whereas the second term takes the form

$$\nabla_x\times\left\{n_y\times\nabla_y\left[\frac{w(y,t^*)}{R}\right]-\frac{1}{R}n_y\times[\nabla_y w(y,\tau)\,]_{\tau=t^*}\right\}.$$

These are combined with expression (A4) and the resulting expression is inserted into the integrand of Eq. (A3). Using Stoke's theorem (over the closed surface $\mathscr{S}$) to eliminate one of the terms, the resulting expression for $\nabla_x Q$ becomes

$$\nabla_x Q = \int_{\mathscr{S}} \frac{H(t^*)}{2\pi R c^2} n_y w_{tt}(y,t^*) d\sigma_y - \nabla_x \times \int_{\mathscr{S}} \frac{H(t^*)}{2\pi R} n_y \times [\nabla_y w(y,\tau)]_{\tau=t^*} d\sigma_y. \tag{A5}$$

Then taking the limit as $x \in G_e \to x \in \mathscr{S}$, it follows from Eqs. (13) and (A5) that

$$\mathfrak{L} w[x,t] = \int_{\mathscr{S}} \frac{H(t^*)}{2\pi R c^2} n_0 \cdot n_y w_{tt}(y,t^*) d\sigma_y - n_0 \cdot \nabla_x \times \int_{\mathscr{S}} \frac{H(t^*)}{2\pi R} n_y \times [\nabla_y w(y,\tau)]_{\tau=t^*} d\sigma_y. \tag{A6}$$

The second term involves the tangential derivatives of a single layer potential that is continuous across $\mathscr{S}$. This can be reduced further by an approach similar to that used by Gunter.[13]

APPENDIX B: EVALUATION OF THE INTEGRAL J

From Eq. (34) the integral J can be expressed in the form

$$J = \int\int \frac{\partial}{\partial \xi_1'} \left[\chi \frac{h_2' h_3'}{h_1'} \left(\frac{\partial 1/R}{\partial \xi_1'} \right) \right] d\xi_2'\, d\xi_3', \tag{B1}$$

where

$$\chi = u\left(\xi_2', \xi_3', t - \frac{R}{c}\right) + \frac{R}{c} u_t \left(\xi_2', \xi_3', t - \frac{R}{c}\right). \tag{B2}$$

The integral expression for J can be reduced to

$$\int\int \left\{ \chi \frac{\partial}{\partial \xi_1'} \left(\frac{h_2' h_3'}{h_1'} \frac{\partial 1/R}{\partial \xi_1'} \right) + \frac{h_2' h_3'}{h_1'} \times \frac{1}{R} \left(\frac{\partial R}{\partial \xi_1'} \right)^2 \frac{1}{c^2} u_{tt} \left(\xi_2', \xi_3', t - \frac{R}{c} \right) \right\} d\xi_2'\, d\xi_3'. \tag{B3}$$

With the point $x \in G_e$, we have $\nabla^2(1/R) = 0$, yielding the relation

$$\frac{\partial}{\partial \xi_1'} \left(\frac{h_2' h_3'}{h_1'} \frac{\partial 1/R}{\partial \xi_1'} \right) = -h_1' h_2' h_3' \nabla_T^2 \left(\frac{1}{R} \right), \tag{B4}$$

where

$$h_1' h_2' h_3' \nabla_T^2 = \frac{\partial}{\partial \xi_2'} \left(\frac{h_1' h_3'}{h_2'} \frac{\partial}{\partial \xi_2'} \right) + \frac{\partial}{\partial \xi_3'} \left(\frac{h_1' h_2'}{h_3'} \frac{\partial}{\partial \xi_3'} \right).$$

Employing expansion (B4), the first integral in the integral (B3) can be integrated by parts twice yielding

$$J = \int\int_{\mathscr{S}} \frac{h_1'}{R} \left\{ -\nabla_T^2 \chi + \left(\frac{1}{h_1'} \frac{\partial R}{\partial \xi_1'} \right)^2 \times \frac{1}{c^2} u_{tt}\left(y, t - \frac{R}{c} \right) \right\} d\sigma_y. \tag{B5}$$

There is no contribution from the end points since the surface $\mathscr{S}$ is closed.

It can be shown by differentiation that

$$\nabla_T^2 \chi = \left\{ \nabla_T^2 u(y,\tau) - \sum_{i=2}^{3} \left(h_{i'} \frac{\partial R}{\partial \xi_i'} \right)^2 \frac{1}{c^2} \frac{\partial^2 u}{\partial \tau^2}(y,\tau) \right\}_{\tau = t - R/c} + \frac{R}{c} \nabla_T^2 u_t \left(\xi_2', \xi_3', t - \frac{R}{c} \right). \tag{B6}$$

Using the relation

$$\sum_{i=1}^{3} \left(\frac{1}{h_i'} \frac{\partial R}{\partial \xi_i'} \right)^2 = |\nabla_y R|^2 = 1, \tag{B7}$$

it immediately follows that expression (B5) reduces to

$$J = -\int\int_{\mathscr{S}} \frac{h_1'}{R} \left\{ \nabla_T^2 u(y,\tau) - \frac{1}{c^2} \frac{\partial^2 u}{\partial \tau^2}(y,\tau) \right\}_{\tau = t - R/c} d\sigma_y. \tag{B8}$$

One can immediately take the limit as $x \in G_e \to x \in \mathscr{S}$, i.e., $\xi_1 \to \xi_1^0$.

APPENDIX C: EVALUATIONS OF KERNELS $k(\rho,t)$, $m(\rho,t)$

The kernel $k(\rho,t)$ has the form [Eq. (65)],

$$k(\rho,t) = \int_{\mathscr{S}} c \frac{\delta(ct - R)}{2\pi R} d\sigma_x,$$

where the integral is over the cylindrical surface $\rho = \text{const}$, with R the distance between two points x and y on the surface. Choose the coordinate system so that y has cylindrical coordinates $(\rho,0,0)$ and x cylindrical coordinates (ρ,θ,x_3), $R^2 = x_3^2 + 2\rho^2(1 - \cos\theta)$. Thus setting $p(\theta) = 2\rho \sin(\theta/2)$, one obtains

$$\begin{aligned} k(\rho,t) &= \frac{2c}{\pi} \int_0^{\pi} \int_0^{\infty} \frac{\delta(ct - R)}{R} \rho\, dx_3\, d\theta \\ &= \frac{2\rho c}{\pi} \int_0^{\pi} \int_{p(\theta)}^{\infty} \frac{\delta(ct - R)}{\sqrt{R^2 - p(\theta)^2}} dR\, d\theta \\ &= \frac{2\rho c}{\pi} \int_0^{\pi} \frac{H(ct - p(\theta))}{\sqrt{(ct)^2 - p(\theta)^2}} d\theta. \end{aligned}$$

Set $\zeta = ct/2\rho$, $\sin(\theta/2) = \zeta\eta$, then

$$k(\rho,t) = \frac{2c}{\pi} \int_0^{1/\zeta} \frac{H(ct(1 - \eta))}{\sqrt{(1 - \eta^2)(1 - \zeta^2\eta^2)}} d\eta.$$

This can be expressed in terms of the complete (Legendre) elliptic integral of the first kind[16]

$$K(\zeta) = \int_0^1 \frac{1}{\sqrt{(1 - \eta^2)(1 - \zeta^2\eta^2)}} d\eta,$$

as follows:

$$k(\rho,t) = \begin{cases} (2c/\pi) K(\zeta), & 0 \leqslant t \leqslant 2\rho/c, \\ (4\rho/\pi t) K(1/\zeta), & 2\rho/c < t. \end{cases} \tag{C1}$$

The evaluation of $m_j(\rho,t), j = 1,2$ from Eq. (75) follows the same way as for $k(\rho,t)$. Note that in cylindrical coordinates the extra factor in the integrand becomes

$$\frac{1}{R^{j-1}} \frac{\partial R}{\partial n_y} = \frac{n_y \cdot (y - x)}{R^j} = \frac{\rho(1 - \cos\theta)}{R^j}.$$

Thus the expressions for $m_j(\rho,t)$ can be reduced to

$$m_2(\rho,t) = \frac{c}{\pi\rho}\int_0^{1/\zeta} \frac{H(ct(1-\eta))\eta^2}{\sqrt{(1-\eta^2)(1-\zeta^2\eta^2)}}\,d\eta, \qquad \text{(C2)}$$

$$m_1(\rho,t) = ctm_2(\rho,t). \qquad \text{(C3)}$$

The integral on the right-hand side of Eq. (C2) is just the elliptic integral[16] $D(\zeta)$ when $\zeta<1$ and $\zeta^{-3}D(1/\zeta)$ when $\zeta>1$. These in turn can be expressed in terms of the elliptic integral K and its derivative, giving

$$m_2(\rho,t) = \frac{c}{\pi\rho}\,[K(\zeta) + (\zeta - 1/\zeta)K'(\zeta)], \quad 0\leqslant t<2\rho/c, \qquad \text{(C4)}$$

$$= \frac{c}{\pi\rho}\,\zeta^{-3}[K(1/\zeta) + (1/\zeta - \zeta)K'(1/\zeta)], \quad 2\rho/c<t. \qquad \text{(C5)}$$

For $0\leqslant t<2\rho/c$, it follows that

$$m(\rho,t) = 2m_2(\rho,t) + t\frac{\partial}{\partial t}m_2(\rho,t)$$
$$= (c/\pi\rho)[(\zeta^2-1)K'' + (4\zeta - 1/\zeta)K' + 2K].$$

Using the second-order differential equation satisfied by K (represented by a hypergeometric function), this reduces to

$$m(\rho,t) = (c/\pi\rho)[K(\zeta) + \zeta K'(\zeta)]. \qquad \text{(C6)}$$

[1] J. P. Corones, M. E. Davison, and R. J. Krueger, "Wave splittings, invariant imbedding and inverse scattering," in *Inverse Optics*, Proc. SPIE 413, edited by A. J. Devaney (SPIE, Bellingham, WA, 1983), pp. 102–106.

[2] M. Davison, "A general approach to splitting and invariant imbedding techniques for linear wave equations," to appear in J. Math. Anal. Appl.

[3] R. Bellman and G. N. Wing, *An Introduction to Invariant Imbedding* (Wiley, New York, 1975).

[4] R. Redheffer, "On the relation of transmission-line theory to scattering and transfer," J. Math. Phys. (Cambridge, MA) **41**, 1 (1962).

[5] J. Corones and R. J. Krueger, "Obtaining scattering kernels using invariant imbedding," J. Math. Anal. Appl. **95**, 393 (1983).

[6] J. P. Corones, M. E. Davison, and R. J. Krueger, "Direct and inverse scattering in the time domain by invariant imbedding techniques," J. Acoust. Soc. Am. **74**, 1535 (1983).

[7] J. P. Corones and R. J. Krueger, "Higher-order parabolic approximations to time-independent wave equations," J. Math. Phys. **24**, 2301 (1983).

[8] L. Fishman and J. J. McCoy, "Derivation and application of extended wave parabolic wave theories, I. The factorized Helmholtz equation," J. Math. Phys. **25**, 285 (1984).

[9] A. E. Yagle and B. L. Levy, "Layer stripping solutions of multidimensional inverse scattering problems," J. Math. Phys. **27**, 1701 (1986).

[10] V. H. Weston, "Factorization of the wave equation in higher dimensions," J. Math. Phys. **28**, 1061 (1987).

[11] S. G. Mikhlin, *Linear Equations of Mathematical Physics* (Holt, Rinehart, and Winston, New York, 1967).

[12] V. S. Vladimirov, *Equations of Mathematical Physics* (Dekker, New York, 1971).

[13] N. M. Günter, *Potential Theory* (Ungar, New York, 1967).

[14] K. Kreider, Ph.D. thesis, Purdue University, 1986.

[15] C. Miranda, *Partial Differential Equations of Elliptic Type* (Springer, Berlin, 1970).

[16] A. Erdelyi, W. Māgnus, E. Oberhettinger, and F. G. Tricomi, *Higher Transcendental Functions* (McGraw-Hill, New York, 1953), Vol. 2.

The Hamiltonian structure of the modified and fifth-order Korteweg–de Vries equation

G. W. Kentwell
Department of Theoretical Physics, Research School of Physical Sciences, Australian National University, Canberra, A.C.T. 2601, Australia

(Received 6 May 1985; accepted for publication 26 August 1987)

Dirac's theory of constraints is used to derive the Hamiltonian structure of the modified and fifth-order Korteweg–de Vries equations. A transformation to "physical" variables is performed on the canonical structure and is shown to be equivalent to the Dirac bracket.

I. INTRODUCTION

Hamiltonian descriptions of nonlinear hydrodynamics have received considerable attention in the literature in recent years.[1,2] This is primarily due to the realization that canonical variables are not essential for a Hamiltonian description. In conventional or canonical Hamiltonian formalisms, the vector space of functions in the Poisson bracket (PB) comprise a Lie algebra. Furthermore, the equations of motion are given by the Liouville equation, which conserves the density in the canonical phase space.

It is, however, possible to construct PB's in terms of physical variables (which are not canonical) which also satisfy the Lie algebra axioms of antisymmetry and the Jacobi identity. The equations of motion are still derived from a Liouville equation in "physical" space. As distinct from the canonical PB, the cosymplectic tensor is no longer constant, but depends on the physical variables.

It is difficult to ascertain who introduced these generalizations first, since it appears that generalized Hamiltonian descriptions of this kind were developed independently by Birkoff,[3] Born and Infeld,[4] Pauli,[5] and Martin.[6] Without doubt, however, the "resurgence" in generalized Hamiltonian descriptions, particularly in field theory, is primarily due to Gardner[7] and Morrison.[8]

Gardner, using a decomposition of the field into normal modes, derived the generalized Poisson bracket (GPB) for the Korteweg–de Vries (KdV) equation, which describes weakly nonlinear dispersive waves for a variety of systems.[9] This then led to interest in the Hamiltonian structure of other integrable nonlinear evolution equations. In particular it provided a basis for the work by Fadeev and Zakharov[10] who show that canonical transformations of the action-angle type, for the Hamiltonian structure, gives the scattering data in the inverse scattering transform method.[11]

Morrison[8] presented the GPB for the Maxwell–Vlasov system, while Morrison and Greene[12] presented the GPB for magnetohydrodynamics. Since this work, most other plasma hydrodynamic models have been cast into GPB form. The various approaches used to derive these GPB's may be found in Ref. 1. There are essentially three approaches. The first[8] proceeds by guesswork but it is clear that the Jacobi identity must be verified. The second[13] proceeds by group theoretic methods which uses the moment mapping of coadjoint group actions of a particular Lie group, which is the configuration space for the system. The third method[14] is via a direct change of representation from a Hamiltonian system to a noncanonical system. In this latter approach, the canonical variables are replaced by the physical variables of the system. We call this approach the direct approach and it provides a straightforward procedure to derive the GPB for systems that possess a canonical Hamiltonian structure. For the KdV equation and its generalizations, this procedure is not so easy, since the Lagrangian for these equations is singular, and thus invalidates the usual Legendre transformation procedure for going from a Lagrangian to a Hamiltonian form. This singular behavior implies $|\partial^2 L/\partial q_i q_j| = 0$, where L is the Lagrangian density and the q_i are the generalized coordinates.

Dirac[15] was the first to consider how canonical Hamiltonian formalisms are modified for singular Lagrangian systems. The basic motivation was for quantization and is very important in the canonical quantization approaches to gravity[16] and relativistic action at a distance theories.[17] The basic idea behind Dirac's mechanics is to introduce multipliers into the Hamiltonian (which includes the usual canonical part). These multipliers imply the existence of constraints among the canonical momenta. By requiring that the constraints be preserved in time, the Lagrange multipliers may be determined and used in the Hamiltonian. However, at this stage, the constraints cannot be used in the PB until equations of motion have been determined. Dirac introduced a new bracket, the Dirac bracket (which is in fact a GPB), that avoids this restriction and reduces the problem to the physically relevant degrees of freedom. Such a construction is necessary for a quantization procedure.

Nutku[18] has used the Dirac procedure to derive the Hamiltonian for the KdV equation. He did not, however, discuss the Hamiltonian structure of the KdV equation. This has recently been done by Lund[19] and by Bergvelt and De Kerf.[20] In this paper we will use the Dirac theory of constraints and derive the Hamiltonian structure for the modified KdV equation and the next member in the KdV hierarchy, the fifth-order KdV equation. This is accomplished in two ways. First, we derive the Hamiltonian structure from the Dirac bracket. Second, as an alternative method, we derived the same results by the direct approach and without the need of the Dirac bracket.

II. DERIVATION OF THE HAMILTONIAN

The modified KdV equation is given by[9]

$$u_t + 6u^2u_x + u_{3x} = 0\,, \tag{1}$$

0022-2488/88/010046-03$02.50

and may be derived from the variational principle

$$\delta I = 0, \quad I = \int \mathscr{L}\, dx\, dt\,. \tag{2}$$

The Lagrangian density $\mathscr{L}$ is given by

$$\mathscr{L} = \tfrac{1}{2}\phi_t\, \phi_x + \tfrac{1}{2}\phi_x^4 - \tfrac{1}{2}\phi_{xx}^2\,, \tag{3}$$

where we have defined $u = \phi_x$. It is clear that the canonical momenta is

$$p_\phi \equiv \frac{\delta L}{\delta \phi_t} = \frac{1}{2}\phi_x\,, \tag{4}$$

so we cannot express the field ϕ in terms of the momenta. Equation (5) is therefore a constraint on the system, so to derive a canonical Hamiltonian structure we have to use the Dirac theory of constraints.[15,21] The Hamiltonian density $\mathscr{H}$ defined by $H = \int \mathscr{H}\, dx$, is

$$\mathscr{H} = p_\phi\, \phi_t - \mathscr{L} = -\tfrac{1}{2}u^4 + \tfrac{1}{2}u_x^2\,. \tag{5}$$

The fifth-order KdV equation is[22]

$$u_t + 30u^2 u_x + 10uu_{3x} + 20u_x\, u_{xx} + u_{5x} = 0\,. \tag{6}$$

This equation may be seen to follow from a Lagrangian density given by

$$\mathscr{L} = \tfrac{1}{2}\phi_t\, \phi_x + \tfrac{5}{2}\phi_x^4 + \tfrac{1}{2}\phi_{3x}^2 - 5\phi_x\, \phi_{xx}^2\,. \tag{7}$$

The canonical momenta is again given by Eq. (4), while the Hamiltonian density is clearly

$$\mathscr{H}' = -\tfrac{5}{2}u^4 + 5uu_x^2 - \tfrac{1}{2}u_{2x}^2\,. \tag{8}$$

III. CONSTRUCTION OF THE DIRAC BRACKET

In the Dirac procedure, every first or second class constraint introduces a Lagrange multiplier χ_i into $\mathscr{H}$ via

$$\mathscr{H}_D = \mathscr{H} + \sum_{i=1}^{r} \chi_i\, C_i\,, \tag{9}$$

where r is the number of constraints while the C_i's are the constraints such as Eq. (4). In our case we have

$$C_1 \equiv p_\phi - \tfrac{1}{2}\phi_x\,. \tag{10}$$

Preservation of constraints imply

$$\dot{C}_i = \{C_i, \mathscr{H}\} = 0\,, \tag{11}$$

where

$$\{A, B\} = \int \left(\frac{\delta A}{\delta \phi}\frac{\delta B}{\delta p_\phi} - \frac{\delta A}{\delta p_\phi}\frac{\delta B}{\delta \phi} \right) dx\,, \tag{12}$$

allows the determination of the χ_i's. To derive Eqs. (1) and (6) from Hamilton's equations requires the use of Eqs. (9) and (12), but we cannot use the constraint equations, such as Eq. (10), until the Poisson bracket operation is performed. On the other hand, to use the simpler canonical Hamiltonian density $\mathscr{H}$, we have to replace Eq. (12) by

$$\{A, B\}^* = \{A, B\} - \int dw\, dv \{A, C_m\} D_{mn}^{-1} \{C_n, B\}\,, \tag{13}$$

where D_{mn}^{-1} is defined via

$$\int dy\, D_{mn}^{-1}(x,y) D_{nt}(y,z) = \delta_{mt}\delta(x-z) \tag{14}$$

and

$$D_{mn} = \{C_m, C_n\}\,. \tag{15}$$

For the modified KdV equation we have

$$D_{11} = -\frac{\partial}{\partial x}\delta(x-y)\,,$$

so Eq. (14) implies

$$D_{11}^{-1} = -\tfrac{1}{2}\varepsilon(w-v)\,, \tag{16}$$

where $\varepsilon(x) \equiv \mathrm{sgn}(x)$. Equation (13) is then

$$\{A(v), B(v)\}^* = \int dx' \frac{\delta A}{\delta u(x')} \frac{\partial}{\partial x} \frac{\delta B}{\delta u(x')}\,, \tag{17}$$

which is the desired noncanonical Poisson bracket. The equation of motion for u is

$$u_t = \{u, \mathscr{H}\}^* = \frac{\partial}{\partial x} \frac{\delta \mathscr{H}}{\partial u(x)}\,, \tag{18}$$

where $\mathscr{H}$ is given by Eq. (5). Since the constraints are the same for the fifth-order KdV equation, Eq. (6) follows from Eq. (18) where $\mathscr{H}$ is now given by Eq. (8).

IV. THE DIRECT APPROACH

Let us recall a few basic notions from Hamiltonian mechanics.[23] If we denote the canonical variables p_i and $q_i\,(i = 1,...,k)$ for some dynamical system by ω^μ, where $\mu = 1,2,...,2k$, then the Poisson bracket may be written as

$$\{A(\omega), B(\omega)\} = \varepsilon^{\mu\nu} \frac{\partial A}{\partial \omega^\mu} \frac{\partial B}{\delta \omega^\nu}\,, \tag{19}$$

where $\varepsilon^{\mu\nu}$ is the antisymmetric tensor. If A and B are transformed into A' and B' by $\omega^\mu \to z^\mu$ then

$$A'(z) = A(\omega), \quad B'(z) = B(\omega)\,,$$

so Eq. (19) becomes

$$\{A(z), B(z)\} = \eta^{\mu\nu}(z) \frac{\partial A}{\partial z^\mu} \frac{\partial B}{\partial z^\nu}\,, \tag{20}$$

where the cosymplectic form $\eta^{\mu\nu}$ is no longer constant and is given by

$$\eta^{\mu\nu}(z) \equiv \{z^\mu, z^\nu\}\,.$$

More generally, for functionals of p and q, Eq. (20) becomes

$$\{A(\alpha), B(\alpha)\} = \int dx\, dx' \frac{\delta A}{\delta \alpha_i(x)}\, O_{ij}\, \frac{\delta B}{\delta \alpha_j(x')} \tag{21}$$

with

$$O_{ij} \equiv \{\alpha_i(x), \alpha_j(x')\}\,. \tag{22}$$

If $A(\alpha) = \alpha_i(x)$, the evolution of the α's is given by

$$\dot{\alpha}_i = \sum_j \int dx'\, O_{ij}(x,x') \frac{\delta \mathscr{H}}{\delta \alpha_j(x')}\,, \tag{23}$$

since $\delta\alpha_i / \delta\alpha_i(r') = \delta(r - r')$. We wish to consider the transformation from $(\phi, p_\phi) \to (\phi_x, p_\phi)$.

The only nonzero element in O_{ij} is $\{\phi_x, p_\phi\}$ and is given by

$$\{\phi_x, p_\phi\} = -\frac{\partial}{\partial x'}\delta(x - x')\,. \tag{24}$$

Using $u = \phi_x$ and $\mathscr{H}$ given by Eq. (5) or Eq. (8), Eq. (23) becomes

$$\dot{u} = \{u,\mathscr{H}\},$$

where

$$\{A,B\} = \int \frac{\delta A}{\delta u}\frac{\partial}{\partial x}\frac{\delta B}{\delta u}\,dx\,,$$

which is just the Dirac bracket we calculated before. It is clear that Eq. (24) is nothing other than D_{11}.

V. CONCLUSION

In this paper we have derived the Hamiltonian structure for the modified KdV equation and the fifth-order KdV equation by two methods. First, the Dirac theory of constraints was used from which we calculated the Dirac bracket. This noncanonical, or generalized, Poisson bracket was shown explicitly to be the bracket that would be obtained by a simple variable change from the more conventional canonical Poisson bracket.

ACKNOWLEDGMENT

I would like to thank the Queen Elizabeth II National Research Fellowship scheme for financial support.

[1] See the various papers in Contemp. Math. **28** (1984).

[2] P. J. Morrison, in *Mathematical Methods in Hydrodynamics and Integrability in Related Dynamical Systems*, AIP Conference Proceedings, La Jolla, December 1981, edited by M. Tabor and Y. Treve (AIP, New York, 1982).

[3] R. M. Santilli, *Foundations of Theoretical Mechanics II-The Birkoffian Generalization of Hamiltonian Mechanics* (Springer, New York, 1981).

[4] M. Born and L. Infeld, Proc. R. Soc. London Ser. A **150**, 141 (1935).

[5] W. Pauli, Nuovo Cimento **10**, 648 (1953).

[6] J. L. Martin, Proc. R. Soc. London Ser. A **251**, 536 (1959).

[7] C. S. Gardner, J. Math. Phys. **12**, 1548 (1971).

[8] P. J. Morrison, Phys. Lett. A **80**, 383 (1980).

[9] A. C. Scott, F. Y. F. Chu, and D. W. McLaughlin, Proc. IEEE **61**, 1443 (1973).

[10] L. Fadeev and V. E. Zakharov, Func. Anal. Appl. **5**, 280 (1972).

[11] M. Ablowitz, D. Kaup, A. Newell, and H. Segur, Phys. Rev. Lett. **19**, 1095 (1967).

[12] P. J. Morrison and J. M. Greene, Phys. Rev. Lett. **45**, 790 (1980); **48**, 569 (E)(1982).

[13] J. E. Marsden and A. Weinstein, Physica D **4**, 394 (1982).

[14] I. Bialynicki-Birula and Z. Iwiniski, Rep. Math. Phys. **4**, 139 (1973).

[15] P. A. M. Dirac, "Lectures on Quantum Mechanics," Belfer Graduate School Monograph Series No. 2, Yeshiva University, 1964.

[16] J. Isenberg and J. Nester, in *General Relativity and Gravitation*, edited by A. Held (Plenum, New York, 1980), Vol. I, p. 23.

[17] See the various papers in Lect. Notes Phys. **162** (1982).

[18] Y. Nutku, J. Math. Phys. **25**, 2007 (1984).

[19] F. Lund, Physica D **18**, 420 (1986).

[20] M. J. Bergvelt and E. A. DeKerf, Lett. Math. Phys. **10**, 13 (1985).

[21] K. Sundermeyer, Lect. Notes Phys. **169** (1982).

[22] M. Ito, J. Phys. Soc. Jpn. **49**, 771 (1980).

[23] E. C. G. Sudarshan and N. Mukunda, *Classical Mechanics* (Wiley, New York, 1974), p. 110.

New integrable nonlinear integrodifferential equations and related solvable finite-dimensional dynamical systems

Y. Matsuno
Department of Physics, Faculty of Liberal Arts, Yamaguchi University, Yamaguchi, 753, Japan

(Received 19 May 1987; accepted for publication 2 September 1987)

A new integrable nonlinear integrodifferential equation (NIDE) is proposed. This equation may be interpreted as a model equation for deep-water waves. The N-periodic and N-soliton solutions for the equation are constructed by means of the bilinear transformation method. These solutions have the same structure as that for the Benjamin–Ono equation which describes internal waves in stratified fluids of great depth. Furthermore, it is shown that the motion of the positions of the poles of solutions is related to certain solvable finite-dimensional dynamical systems described by first-order nonlinear ordinary differential equations. The discussion is also made on a more general NIDE that may be interpreted as a model equation describing nonlinear waves in fluids of finite depth.

I. INTRODUCTION

Recently, much attention has been paid to integrable nonlinear integrodifferential equations (NIDE's) of both physical and mathematical interests such as the Benjamin–Ono (BO) equation,[1–5] the intermediate long wave (ILW) equation,[6,8] the sine–Hilbert equation,[9–11] and some other related NIDE's.[9,12,13] In this paper, we shall propose a new NIDE that exhibits exact N-periodic wave and N-soliton solutions. The equation that we consider here reads

$$u_t - Hu_{tx} - uu_t + u_x \int_x^\infty u_t\,dx + u_x = 0, \quad u = u(x,t), \tag{1.1a}$$

with

$$Hu(x,t) = \frac{1}{\pi} P \int_{-\infty}^{\infty} \frac{u(y,t)}{y-x}\,dy, \tag{1.1b}$$

where the operator H is the Hilbert transform [the symbol P in (1.1b) stands for the Cauchy principal value] and the abbreviations $u_t = \partial u/\partial t$, $u_x = \partial u/\partial x$, and $u_{tx} = \partial^2 u/\partial t\,\partial x$ have been used. Equation (1.1) includes both the definite and indefinite integrals and in this respect Eq. (1.1) is quite different from the known NIDE's mentioned above. We note that Eq. (1.1) is reduced to the following model equation for shallow water waves introduced by Hirota and Satsuma[14]:

$$u_t - u_{txx} - uu_t + u_x \int_x^\infty u_t\,dx + u_x = 0, \tag{1.2}$$

provided that the H operator is replaced formally by an x derivative. Mathematically, this formal derivation is entirely analogous to that of the Korteweg–de Vries (KdV) equation from the BO equation. Physically, a new NIDE (1.1) may be interpreted as a model equation that describes nonlinear waves in fluids of great depth.

In Sec. II, we analyze Eq. (1.1) by means of the well-known bilinear transformation method[15,16] and construct the N-periodic wave and N-soliton solutions. The latter solutions stem quite naturally from the long-wave limit of the former solutions. The initial value problem for a linearized version of Eq. (1.1) is also solved exactly in the last part of this section. In Sec. III, it is shown that the motion of the poles of solutions presented in Sec. II is closely related to certain solvable finite-dimensional dynamical systems described by first-order ordinary differential equations. It then follows from the integrability of Eq. (1.1) that solutions for the dynamical systems are determined by solving *algebraic equations* of order N. This remarkable fact implies an aspect of the integrability of the dynamical systems related to Eq. (1.1). In Sec. IV, we generalize Eq. (1.1) to a more general NIDE that is reduced to Eq. (1.1) and Eq. (1.2) in the deep-water and shallow-water limits, respectively. This equation may describe relevantly nonlinear waves in fluids of finite depth. The N-soliton and some rational solutions for the generalized NIDE are presented and subsequently we show that the NIDE is related to a solvable finite-dimensional dynamical system. In addition, the two limiting procedures, namely the deep-water and shallow-water limits, are taken for both solutions and a dynamical system obtained here. The results are consistent with corresponding solutions and a related dynamical system for Eq. (1.1), in the deep-water limit and those for Eq. (1.2), in the shallow-water limit, respectively. Section V is devoted to the conclusion.

II. EXACT SOLUTIONS

A. *N*-periodic wave solution

First, we focus our attention on a real and finite periodic-wave solution of Eq. (1.1) and seek it in the form

$$u = i\frac{\partial}{\partial x}\ln\left(\frac{f_+}{f_-}\right), \quad f_\pm = f_\pm(x,t), \tag{2.1}$$

where $f_+(f_-)$ is a complex analytic function with zeros lying only in the lower (upper) half complex x plane. The dependent variable transformation (2.1) is the same as that used for the BO[1,2] and the ILW[6,7] equations. It then follows by using the property of the H operator that

$$Hu = -\frac{\partial}{\partial x}\ln(f_+f_-). \tag{2.2}$$

Now, substituting (2.1) and (2.2) into Eq. (1.1) and integrating once with respect to x, Eq. (1.1) is transformed into the following bilinear equation for f_+ and f_-:

$$(iD_t + D_t D_x + iD_x) f_+ \cdot f_- = 0, \tag{2.3}$$

where the integration constant has been taken to be zero and the bilinear operators D_t and D_x are defined by the relation

$$D_t^n D_x^m f_+ \cdot f_- = \left(\frac{\partial}{\partial t} - \frac{\partial}{\partial t'}\right)^n \left(\frac{\partial}{\partial x} - \frac{\partial}{\partial x'}\right)^m \times f_+(x,t) f_-(x',t') \Big|_{\substack{x'=x \\ t'=t}} \quad (n,m = 0,1,...). \tag{2.4}$$

Applying the standard procedure in the bilinear transformation method[15,16] to Eq. (2.3), we have found the following N-periodic wave solution of Eq. (1.1):

$$u = -\sum_{j=1}^{N} k_j + i\frac{\partial}{\partial x}\ln\left(\frac{f_+}{f_-}\right), \tag{2.5a}$$

$$f_- = \sum_{\mu=0,1} \exp\left[\sum_{j=1}^{N} \mu_j(i\xi_j + \phi_j) + \sum_{j<l}^{(N)} \mu_j\mu_l A_{jl}\right], \tag{2.5b}$$

$$f_+ = f_-^* \quad (*\text{: complex conjugate}), \tag{2.5c}$$

$$\xi_j = k_j(x - a_j t - x_{0j}) + \xi_j^{(0)} \quad (j = 1,2,...,N), \tag{2.5d}$$

$$a_j = (1 - k_j \coth\phi_j)^{-1}, \quad \phi_j/k_j > 0 \quad (j = 1,2,...,N), \tag{2.5e}$$

$$\exp A_{jl} = \frac{(a_j - a_l)^2 - a_j a_l (a_j k_j - a_l k_l)(k_j - k_l)}{(a_j - a_l)^2 - a_j a_l (a_j k_j + a_l k_l)(k_j + k_l)}. \tag{2.5f}$$

Here, $\Sigma_{\mu=0,1}$ denotes the summation over all possible combinations of $\mu_1 = 0,1, \mu_2 = 0,1,..., \mu_N = 0,1$, $\Sigma_{j<l}^{(N)}$ means the summation under the condition $j<l$ and k_j, ϕ_j, x_{0j}, and $\xi_j^{(0)}$ $(j = 1,2,...,N)$ are real constants.

For $N = 1$, the solution (2.5) is written explicitly in the form

$$u = \frac{k_1 \tanh\phi_1}{1 + \operatorname{sech}\phi_1 \cos\xi_1} \quad \left(\frac{\phi_1}{k_1} > 0\right), \tag{2.6}$$

which represents a real and finite one-periodic wave solution of Eq. (1.1). Except for the phase velocity a_1, the functional form of (2.6) coincides with the periodic solution of the BO equation presented by Benjamin[17] and Ono.[18] Note that in the BO case, $a_1 = k_1 \coth\phi_1$.

B. *N*-soliton solution

The N-soliton solution is easily constructed by taking the long-wave limit of the N-periodic wave solution (2.5). To show this, we set in (2.5d),

$$\xi_j^{(0)} = \pi \quad (j = 1,2,...,N), \tag{2.7}$$

and take the long-wave limit $k_j \to 0$ $(j = 1,2,...,N)$ with the phase velocities a_j $(j = 1,2,...,N)$ keeping finite values. It then turns out that

$$A_{jl} = \frac{2(a_j + a_l)a_j a_l}{(a_j - a_l)^2} k_j k_l + O(k_j^4). \tag{2.8}$$

Introducing the expansion (2.8) into (2.5b), one finds, in the long-wave limit, the explicit expression of the N-soliton solution of Eq. (1.1) as follows:

$$u = i\frac{\partial}{\partial x}\ln\left(\frac{f^*}{f}\right), \tag{2.9a}$$

$$f = \det M. \tag{2.9b}$$

Here, M is an $N \times N$ matrix whose elements are given by

$$M_{jk} = \begin{cases} i(x - a_j t - x_{0j}) + a_j/(a_j - 1), & \text{for } j = k, \\ [2(a_j + a_k)a_j a_k]^{1/2}/(a_j - a_k), & \text{for } j \neq k, \end{cases} \tag{2.9c}$$

and the phase velocities are restricted by the conditions $a_j > 1$ and $a_j \neq a_k$ for $j \neq k$ $(j, k = 1,2,...,N)$. It is interesting to note that the N-soliton solution of the BO equation

$$u_t + 2uu_x + Hu_{xx} = 0, \tag{2.10}$$

is expressed in the form[1,16]

$$u = i\frac{\partial}{\partial x}\ln\left(\frac{\tilde f^*}{\tilde f}\right), \tag{2.11a}$$

$$\tilde f = \det \tilde M, \tag{2.11b}$$

with an $N \times N$ matrix $\tilde M$ given by

$$\tilde M_{jk} = \begin{cases} i(x - \tilde a_j t - \tilde x_{0j}) + 1/\tilde a_j, & \text{for } j = k, \\ 2/(\tilde a_j - \tilde a_k), & \text{for } j \neq k, \end{cases} \tag{2.11c}$$

where $\tilde a_j$ $(j = 1,2,...,N)$ are positive constants such that $\tilde a_j \neq \tilde a_k$ for $j \neq k$ and $\tilde x_{0j}$ $(j = 1,2,...,N)$ are arbitrary phase constants. Therefore we see that the N-soliton solution of Eq. (1.1) has the same structure as that of the BO equation.

The one-soliton solution is readily derived from (2.9) with $N = 1$. It takes a Lorentzian profile as

$$u = \frac{2b_1}{(x - a_1 t - x_{01})^2 + b_1^2} \quad \left(b_1 = \frac{a_1}{a_1 - 1},\ a_1 > 1\right). \tag{2.12}$$

The amplitude and the velocity of the soliton (2.12) are given, respectively, by $2(a_1 - 1)/a_1$ and a_1. Hence one can observe that the amplitude approaches a constant value 2 indefinitely when the velocity becomes large while it approaches zero in the limit of $a_1 \to 1$. Asymptotic behavior of the N-soliton solution (2.9) for large time is easily obtained following the same argument as that for the BO case.[1,16] The result is expressed simply as a superposition of N independent algebraic solitons as follows:

$$u \underset{t \to \pm\infty}{\sim} \sum_{j=1}^{N} \frac{2b_j}{(x - a_j t - x_{0j})^2 + b_j^2} \quad [b_j = a_j/(a_j - 1),\ a_j > 1]. \tag{2.13}$$

This asymptotic expression shows that no phase shift appears as the result of collisions of solitons in contrast to collisions that take place between the KdV solitons. Thus we have presented the second example of the one space-dimensional algebraic N-soliton solution that is real and finite over all x and t. The first example is, of course, that of the BO equation.[1]

C. Solution for a linearized equation

Here, we consider the initial value problem for a linearized version of Eq. (1.1), namely

$$u_t - Hu_{tx} + u_x = 0, \tag{2.14}$$

with the boundary condition $u(x,t) \to 0$ as $|x| \to \infty$. If $u(x,t)$ is represented in the form of the Fourier transform

$$u(x,t) = \int_{-\infty}^{\infty} v(k)\exp[i(kx - \omega t)]dk, \tag{2.15}$$

we obtain the dispersion relation

$$\omega = k/(1+|k|), \tag{2.16}$$

with the aid of the formula

$$\frac{1}{\pi} P \int_{-\infty}^{\infty} \frac{e^{ikx}}{x} dx = i \frac{k}{|k|}. \tag{2.17}$$

It is interesting to note that for small k, (2.16) behaves like

$$\omega = k - k|k| + O(k^3). \tag{2.18}$$

If we retain only up to the second term in the expansion, the expression (2.18) coincides perfectly with the dispersion relation of the following linearized BO equation

$$u_t + u_x + Hu_{xx} = 0. \tag{2.19}$$

This fact may suggest the suitability for interpreting Eq. (1.1) as a model equation which describes nonlinear wave propagations in fluids of great depth. In comparison with the dispersion relation of Eq. (2.19), Eq. (2.16) is well behaved for a wide range of the values of k, in particular for large k and hence Eq. (1.1) may be more relevant than the BO equation itself as a model equation for deep-water waves.

Now, the unknown function $v(k)$ appeared in (2.15) is determined from the initial value $u(x,0)$ as

$$v(k) = \frac{1}{2\pi} \int_{-\infty}^{\infty} u(x,0) e^{-ikx} dx. \tag{2.20}$$

Substituting (2.16) and (2.20) into (2.15), we obtain a general solution of Eq. (2.14) as follows:

$$u(x,t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} u(y,0) \times \exp\{i[k(x-y) - (1+|k|)^{-1}kt]\} dk\, dy. \tag{2.21}$$

To investigate asymptotic behaviors of (2.21) for large time is an interesting problem. But we shall not be concerned with this problem here and the details will be reported elsewhere.

III. DYNAMICAL SYSTEMS RELATED TO EQ. (1.1)

In this section, we consider the dynamical systems related to Eq. (1.1). The relationships between integrable nonlinear evolution equations and solvable finite-dimensional dynamical systems have been studied extensively by many authors.[19-23] The basic idea due to Kruskal[19] is to investigate the time evolution of the positions of the poles of solutions of nonlinear evolution equations. In the following, we show that Eq. (1.1) is related to certain solvable finite-dimensional dynamical systems. The periodic and nonperiodic dynamical systems are treated separately.

A. Periodic dynamical system

We first consider the periodic dynamical system. As easily seen from (2.5a)–(2.5c), it is possible to express the periodic wave solution (2.5) with $k_j = k$ $(j = 1,2,...,N)$ in the form

$$u = i \frac{\partial}{\partial x} \ln\left(\frac{\tilde{f}_+}{\tilde{f}_-}\right), \tag{3.1a}$$

$$\tilde{f}_- = \prod_{j=1}^{N} \frac{2}{k} \sin\left[\frac{k}{2}(x - x_j)\right], \quad x_j = x_j(t), \tag{3.1b}$$

$$\tilde{f}_+ = \tilde{f}_-^*, \tag{3.1c}$$

where x_j $(j = 1,2,...,N)$ are complex functions of t whose imaginary parts are all positive, i.e., $\operatorname{Im} x_j > 0$ and $x_j \neq x_k$ for $j \neq k$ $(j,k = 1,2,...,N)$. The functions x_j represent positions of the complex poles of the periodic-wave solution (2.5) with $k_j = k$ $(j = 1,2,...,N)$. In order to find the time evolution of x_j, we substitute (3.1) into Eq. (2.3), use the trigonometric identity

$$\cot A \cot B = -1 - (\cot A - \cot B)\cot(A - B), \tag{3.2}$$

and then equate the coefficient of $\cot[k(x - x_j)/2]$ zero. The resultant equations for x_j are written in the form

$$\dot{x}_j = 1 - i\frac{k}{2} \sum_{\substack{l=1 \\ (l \neq j)}}^{N} (\dot{x}_j + \dot{x}_l) \cot\left[\frac{k(x_j - x_l)}{2}\right] + i\frac{k}{2} \sum_{l=1}^{N} (\dot{x}_j + \dot{x}_l^*) \cot\left[\frac{k(x_j - x_l^*)}{2}\right]$$
$$(j = 1,2,...,N), \tag{3.3}$$

where the dot appended to x_j and x_l means the time differentiation. One can also obtain from the coefficient of $\cot[k(x - x_j^*)/2]$ the complex conjugate expressions of Eqs. (3.3). For $N = 1$, Eq. (3.3) reads

$$\dot{x}_1 = 1 + k(\operatorname{Re} \dot{x}_1)\coth(k \operatorname{Im} x_1), \tag{3.4}$$

and this equation is readily integrated to yield the solution

$$x_1 = (1 - k \coth \phi_1)^{-1} t + x_{01} + i\phi_1/k. \tag{3.5}$$

Substituting (3.5) into (3.1), we recover the one-periodic wave solution (2.6). For general N, on the other hand, Eqs. (3.3) and their complex conjugate expressions constitute the system of $2N$ algebraic equations for $\dot{x}_j$ and $\dot{x}_j^*$ $(j = 1,2,...,N)$ and hence it is possible by using Cramer's formula to express these variables in terms of x_n and x_n^* $(n = 1,2,...,N)$ in the form

$$\dot{x}_j = F_j \quad (j = 1,2,...,N), \tag{3.6}$$

together with their complex conjugate expressions, where the F_j are uniquely determined functions of x_n and x_n^* $(n = 1,2,...,N)$. The explicit functional forms of F_j will not be written here. The system of equations (3.6) consists of quite complicated first-order nonlinear ordinary differential equations and hence it could not be solved analytically in general. Nevertheless, in the present situation, one can obtain explicit periodic-wave solutions of Eqs. (3.6). In order to clarify this statement, we compare the expressions (2.5a)–(2.5c) with $k_j = k$ $(j = 1,2,...,N)$ and the expressions (3.1a)–(3.1c). It then follows from (2.5b) with $k_j = k$ that

$$f_- = c_0 e^{ikNx} + c_1 e^{ik(N-1)x} + \cdots + 1$$
$$= c_0[z^N + (c_1/c_0)z^{N-1} + \cdots + c_0^{-1}] \quad (z = e^{ikx}), \tag{3.7}$$

where c_j $(j = 1,2,...,N)$ are known functions expressible in terms of t and various constant parameters. Consequently, the $\exp(ikx_j)$ are determined by solving the *algebraic equation* of order N, $f_- = 0$, with f_- being given by (3.7). In other words, this result reveals an aspect of the complete integrability of the system of equations (3.6).

B. Nonperiodic dynamical system

Next, we investigate the nonperiodic dynamical system. The time evolutions for x_j $(j = 1,2,...,N)$ are derived quite naturally by taking the long-wave limit, $k \to 0$ in Eqs. (3.3). The equations corresponding to Eqs. (3.3) are written in the form

$$\dot{x}_j = 1 - i \sum_{\substack{l=1 \\ (l \neq j)}}^{N} \frac{\dot{x}_j + \dot{x}_l}{x_j - x_l} + i \sum_{l=1}^{N} \frac{\dot{x}_j + \dot{x}_l^*}{x_j - x_l^*}$$

$$(j = 1,2,...,N). \tag{3.8}$$

The solutions for this system of equations are readily found by solving the *algebraic equation* of order N, $f = 0$, where f is given by (2.9b) with (2.9c). Finally, it should be remarked that the dynamical system related to the BO equation (2.10) is completely integrable and it is expressed in the form[3,4]

$$\dot{x}_j = -2i \sum_{\substack{l=1 \\ (l \neq j)}}^{N} \frac{1}{x_j - x_l} + 2i \sum_{l=1}^{N} \frac{1}{x_j - x_l^*}$$

$$(j = 1,2,...,N). \tag{3.9}$$

IV. GENERALIZATION TO MORE GENERAL NIDE

In this section, we generalize Eq. (1.1) to a more general NIDE that is reduced to Eq. (1.1) in the deep-water limit and to Eq. (1.2) in the shallow-water limit, respectively, and construct the N-soliton solution together with some rational solutions for the NIDE. We also investigate the motion of the poles of the N-soliton solution to show that the generalized NIDE is related to certain solvable finite-dimensional dynamical systems. Since the discussion is almost the same as that for Eq. (1.1), we shall not enter into detail but present only the main results.

A generalized version of Eq. (1.1) which we propose here reads

$$u_t - Tu_{tx} - uu_t + u_x \int_x^{\infty} u_t \, dx + u_x = 0, \quad u = u(x,t), \tag{4.1a}$$

with the operator T defined by

$$Tu(x,t) = \frac{1}{2\delta} P \int_{-\infty}^{\infty} \left\{ \coth\left[\frac{\pi(y-x)}{2\delta}\right] - \operatorname{sgn}(y-x) \right\} u(y,t) dy, \tag{4.1b}$$

where δ is a positive parameter that may be interpreted as a depth of fluids. The T operator has been first introduced by Joseph[24–26] in the context of his NIDE which describes nonlinear waves in stratified fluids of finite depth. Presently, his equation is known as the ILW equation.[6–8] In the deep-water limit $\delta \to \infty$ the T operator is reduced to the H operator defined by (1.1b) while in the shallow-water limit $\delta \to 0$ it takes the form

$$Tu = \delta u_x/3 + \delta^3 u_{xxx}/45 + O(\delta^5). \tag{4.2}$$

Therefore Eq. (4.1) is an intermediate version between Eq. (1.1) and Eq. (1.2).

A. N-soliton solution

First, introduce the following dependent variable transformation:

$$u = i \frac{\partial}{\partial x} \ln\left(\frac{f_+}{f_-}\right), \tag{4.3a}$$

with

$$f_+(x,t) = f(x - i\delta, t), \tag{4.3b}$$

$$f_-(x,t) = f(x + i\delta, t), \tag{4.3c}$$

where the complex function $f(z,t)$ is such that $f(z - i\delta, t)$ has no zero in the region $0 \leqslant \operatorname{Im} z \leqslant 2\delta$. Then, it is straightforward by using the Cauchy residue theorem to show that

$$Tu_x = -\frac{\partial^2}{\partial x^2} \ln(f_+ f_-) + \delta^{-1} u. \tag{4.4}$$

Substituting (4.3a) and (4.4) into Eq. (4.1) and integrating once with respect to x, we obtain the bilinear form of Eq. (4.1) as follows:

$$[i(1 - \delta^{-1}) D_t + D_t D_x + i D_x] f_+ \cdot f_- = 0. \tag{4.5}$$

In comparison with Eq. (2.3), Eq. (4.5) differs only from a numerical factor in front of the operator D_t and therefore it may share many of the integrability properties of Eq. (2.3).

The procedure for constructing the N-soliton solution of Eq. (4.5) is a routine work in the context of the bilinear formalism.[15,16] The result is written compactly as follows:

$$f_- = f_+^* = \sum_{\mu=0,1} \exp\left[\sum_{n=1}^{N} \mu_n (\delta^{-1} \gamma_n \theta_n + i\gamma_n) + \sum_{j<m}^{(N)} \mu_j \mu_m B_{jm}\right], \tag{4.6a}$$

with

$$\theta_n = x - a_n t - x_{0n} \quad (n = 1,2,...,N), \tag{4.6b}$$

$$a_n = (1 - \delta^{-1} + \delta^{-1} \gamma_n \cot \gamma_n)^{-1}, \quad 0 < \gamma_n < \pi \quad (n = 1,2,...,N), \tag{4.6c}$$

$$\exp B_{jm} = \frac{\delta^2 (1 - \delta^{-1})(a_j - a_m)^2 + a_j a_m (a_j \gamma_j - a_m \gamma_m)(\gamma_j - \gamma_m)}{\delta^2 (1 - \delta^{-1})(a_j - a_m)^2 + a_j a_m (a_j \gamma_j + a_m \gamma_m)(\gamma_j + \gamma_m)}, \tag{4.6d}$$

where γ_n and x_{0n} $(n = 1,2,...,N)$ are constants. The explicit one-soliton solution follows from (4.3a) and (4.6) with $N = 1$ as

$$u = \frac{\delta^{-1} \gamma_1 \sin \gamma_1}{\cosh[\delta^{-1} \gamma_1 (x - a_1 t - x_{01})] + \cos \gamma_1}, \tag{4.7}$$

which has the same functional form as that of the one-soliton solution of the ILW equation.[6-8,24] The interaction between solitons is easily investigated by using the explicit formula (4.6) for the N-soliton solution. The asymptotic form of the N-soliton solution is simply expressed as a superposition of N independent one-soliton solutions (4.7). In this case, however, the phase shift appears as the result of collisions between solitons. Since the explicit formula for the phase shift is easily derived following a procedure similar to that for the ILW equation,[6] the result will not be presented here.

We can also obtain more general periodic solutions of Eq. (4.1) which are expressed in terms of Riemann's theta function on the basis of the bilinear equation (4.5). The procedure for constructing solutions is almost the same as that for the ILW equation.[16,27] Details will not be discussed here.

Now, we consider the deep- and shallow-water limits, respectively, of the bilinear equation (4.5) and the N-soliton solution (4.6).

1. Deep-water limit: $\delta \to \infty$

In this limit, Eq. (4.5) is reduced to Eq. (2.3) as expected. For the purpose of a limiting procedure for the N-soliton solution (4.6), it is appropriate to introduce the positive constants b_n through the relations

$$\gamma_n = \pi(1 - b_n/\delta) \quad (n = 1,2,...,N). \tag{4.8}$$

Then, it follows in the deep-water limit $\delta \to \infty$ that

$$\cot \gamma_n = -\delta/\pi b_n + \pi b_n/3\delta + O(\delta^{-3}), \tag{4.9a}$$

$$a_n = b_n/(b_n - 1) \quad (b_n > 1), \tag{4.9b}$$

$$B_{jm} = -\frac{2(a_j + a_m)a_j a_m}{(a_j - a_m)^2}\left(\frac{\pi}{\delta}\right)^2 + O(\delta^{-4}). \tag{4.9c}$$

Substituting (4.8) and (4.9) into (4.3) and (4.6), one finds that the N-soliton solution coincides perfectly with that of Eq. (1.1), namely the expression (2.9). The one-soliton solution (4.7) is of course reduced to (2.12), the one-soliton solution of Eq. (1.1).

2. Shallow-water limit: $\delta \to 0$

In this limit, it is appropriate to introduce the variables $\tilde{t}$ and $\tilde{x}$ by

$$t = (\delta/3)^{1/2}\tilde{t}, \tag{4.10a}$$

$$x = (\delta/3)^{1/2}\tilde{x}. \tag{4.10b}$$

Then, it is easy to show by using the properties of the bilinear operator,[16]

$$\exp[-i\delta D_x] f(x)\cdot f(x) = f(x - i\delta) f(x + i\delta), \tag{4.11a}$$

$$D_t D_x^{2m} f\cdot f = 0 \quad (m = 0,1,2,...), \tag{4.11b}$$

that

$$\begin{aligned} D_t f_+\cdot f_- &= D_t \exp(-i\delta D_x) f\cdot f \\ &= -3i D_{\tilde{t}} D_{\tilde{x}} f\cdot f + \tfrac{3}{2} i\delta D_{\tilde{t}} D_{\tilde{x}}^3 f\cdot f + O(\delta^2), \end{aligned} \tag{4.12}$$

$$D_x D_t f_+\cdot f_- = 3\delta^{-1} D_{\tilde{t}} D_{\tilde{x}} f\cdot f - \tfrac{3}{2} D_{\tilde{t}} D_{\tilde{x}}^3 f\cdot f + O(\delta), \tag{4.13}$$

$$D_x f_+\cdot f_- = -3i D_{\tilde{x}}^2 f\cdot f + O(\delta). \tag{4.14}$$

We then have, by substituting (4.10) and(4.12)–(4.14) into Eq. (4.5), the following bilinear equation for f:

$$D_{\tilde{x}}(D_{\tilde{t}} - D_{\tilde{t}} D_{\tilde{x}}^2 + D_{\tilde{x}}) f\cdot f = 0. \tag{4.15}$$

The dependent variable transformation for u follows from (4.3) and (4.10) with the aid of the expressions for small δ,

$$f_+ = f - i\delta f_x + O(\delta^2), \tag{4.16a}$$

$$f_- = f + i\delta f_x + O(\delta^2), \tag{4.16b}$$

as

$$u = 6\frac{\partial^2}{\partial\tilde{x}^2}\ln f. \tag{4.17}$$

Equation (4.15) with (4.17) is nothing but the bilinear form of Eq. (1.2) with the variables $\tilde{t}$ and $\tilde{x}$ instead of t and x, respectively.[14]

In order to derive the explicit functional form for f, we introduce the positive constants p_n by the relations

$$\gamma_n = (3\delta)^{1/2} p_n \quad (n = 1,2,...,N). \tag{4.18}$$

It then turns out that

$$\cos\gamma_n = [(3\delta)^{1/2} p_n]^{-1} - (3\delta)^{1/2} p_n/3 + O(\delta^{3/2}), \tag{4.19a}$$

$$a_n = 1/(1 - p_n^2), \tag{4.19b}$$

$$B_{jm} \equiv \widetilde{B}_{jm} = \frac{(p_j - p_m)^2(-3 + p_j^2 - p_j p_m + p_m^2)}{(p_j + p_m)^2(-3 + p_j^2 + p_j p_m + p_m^2)}, \tag{4.19c}$$

and

$$f = \sum_{\mu=0,1} \exp\left[\sum_{n=1}^{N} \mu_n p_n(\tilde{x} - a_n\tilde{t} - \tilde{x}_{0n}) + \sum_{j<m}^{(N)} \mu_j\mu_m\widetilde{B}_{jm}\right], \quad [\tilde{x}_{0n} = (3/\delta)^{1/2} x_{0n}]. \tag{4.19d}$$

The expression (4.19d) coincides perfectly with that given by Hirota and Satsuma.[14]

B. Rational solutions

Rational solutions of certain nonlinear evolution equations may be constructed by taking an appropriate limit on soliton solutions.[28,29] Owing to the freedom to choose an arbitrary constant, x_{0n} in the present case, it is possible to reduce soliton solutions to corresponding rational ones. In this subsection, we shall briefly discuss some rational solutions that are reduced from the one-soliton solution of Eq. (4.1), namely the expression (4.6) with $N = 1$. The rational solutions reduced from the general N-soliton solution will be presented elsewhere.

Now, it follows from (4.6) with $N = 1$ that the one-soliton solution of Eq. (4.1) is written in terms of the bilinear variables as

$$f = 1 + \exp[\delta^{-1}\gamma_1(x - a_1 t - x_{01})], \tag{4.20a}$$

with

$$a_1 = (1 - \delta^{-1} + \delta^{-1}\gamma_1\cot\gamma_1), \quad 0 < \gamma_1 < \pi. \tag{4.20b}$$

The deduction of the rational solution from the one-soliton solution (4.20) is possible if we choose the phase constant x_{01} as

$$x_{01} = -\pi\delta\gamma^{-1} - i\alpha, \quad |\alpha| > \delta, \tag{4.21}$$

with α being a real constant, and then take a limit $\gamma_1 \to 0$. It should be noted that the condition $|\alpha| > \delta$ is necessary because of the assumption in deriving (4.4). Indeed, a_1 becomes in this limit

$$a_1 = 1 + (3\delta)^{-1}\gamma_1^2 + O(\gamma_1^4), \tag{4.22}$$

and hence f has an expansion for small γ_1,

$$f = -\delta^{-1}\gamma_1(x - t + i\alpha) + O(\gamma_1^2). \tag{4.23}$$

Substituting (4.23) into (4.3) and taking a limit $\gamma_1 \to 0$, we have a rational solution of Eq. (1.1) as follows:

$$u = -2\delta/[(x - t + i\alpha)^2 + \delta^2]. \tag{4.24}$$

The solution (4.24) is regular but complex for real t and x. We now consider the two limiting cases of $\delta \to \infty$ and $\delta = 0$.

1. Deep-water limit: $\delta \to \infty$

In this limit, it is convenient to start from the expression (4.20). Various limiting procedures are possible, which we shall treat separately below.

(a) $\gamma_1 = \pi(1 - b_1/\delta), \quad b_1 > 1.$

In this case, it follows that

$$a_1 = b_1/(b_1 - 1), \tag{4.25a}$$

$$f_- = 1 + \exp[\delta^{-1}\gamma_1(x - a_1 t + i\delta - x_{01})]$$

$$= -\delta^{-1}\pi(x - a_1 t - x_{01} - ib_1) + O(\delta^{-2}), \tag{4.25b}$$

$$f_+ = f_-^*, \tag{4.25c}$$

so that

$$u = i\frac{\partial}{\partial x}\ln\left(\frac{f_+}{f_-}\right)$$

$$= \frac{2b_1}{(x - a_1 t - x_{01})^2 + b_1^2} \quad \left(b_1 = \frac{a_1}{a_1 - 1}\right), \tag{4.25d}$$

which is nothing but the rational one-soliton solution of Eq. (1.1) already given by (2.12).

(b) $\gamma_1 = \pi(1 - c_1/\delta)/2 \ (c_1 > 0), \quad x_{01} = \pi i\delta\gamma_1^{-1}.$

In this case, one finds that

$$a_1 = 1 + O(\delta^{-1}), \tag{4.26a}$$

$$f_- = \pi(x - t - ic_1)/2\delta + O(\delta^{-2}), \tag{4.26b}$$

$$f_+ = 2 + O(\delta^{-1}), \tag{4.26c}$$

so that

$$u = -i/(x - t - ic_1), \tag{4.26d}$$

which is a single pole solution of Eq. (1.1).

(c) $\gamma_1 = \beta(1 - c_1/\delta), \quad \beta \neq \pi/2, \pi, \quad x_{01} = \pi i\delta\gamma_1^{-1}.$

In this case, following the same procedure as case (b), we find a single pole solution (4.26d).

2. Shallow-water limit: $\delta \to 0$

In this limit, we also take $\alpha \to 0$. Then, introduction of the new variables $\tilde{t}$ and $\tilde{x}$ defined by (4.10a) and (4.10b), respectively, into (4.24) yields

$$u = -6/(\tilde{x} - \tilde{t})^2, \tag{4.27}$$

which is a rational solution of Eq. (1.2) with the variables $\tilde{t}$ and $\tilde{x}$ instead of t and x, respectively. This fact can also easily be confirmed by direct substitution of (4.27) into Eq. (1.2).

C. Solution for a linearized equation

An appropriate linearized version of Eq. (4.1) may be written as

$$u_t - Tu_{tx} + u_x = 0. \tag{4.28}$$

The solution of the initial value problem for Eq. (4.28) with the boundary condition $u \to 0$ as $|x| \to \infty$ can easily be constructed by employing the Fourier transform. The result is expressed in the form

$$u(x,t) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} u(y,0)$$

$$\times \exp\{i[k(x - y) - \omega(k)t]\}dk\,dy. \tag{4.29a}$$

Here, $\omega(k)$ is a dispersion relation for Eq. (4.28) given by

$$\omega(k) = k/[1 - \delta^{-1} + k\coth(k\delta)]. \tag{4.29b}$$

In the deep-water limit $\delta \to \infty$, (4.29) is reduced to

$$\omega(k) = k/(1 + |k|), \tag{4.30}$$

which is in accordance with (2.16), the dispersion relation for Eq. (2.14). In the shallow-water limit $\delta \to 0$, on the other hand, it becomes

$$\tilde{\omega}(\tilde{k}) = 1/(1 + \tilde{k}^2), \tag{4.31a}$$

with the new variables $\tilde{\omega}$ and $\tilde{k}$ defined by

$$\tilde{\omega} = (\delta/3)^{1/2}\omega, \tag{4.31b}$$

$$\tilde{k} = (\delta/3)^{1/2}k. \tag{4.31c}$$

The expression (4.31) is just the dispersion relation for the linearized version in Eq. (1.2) with the variables $\tilde{t}$ and $\tilde{x}$ instead of t and x, namely,

$$u_{\tilde{t}} - u_{\tilde{t}\tilde{x}\tilde{x}} + u_{\tilde{x}} = 0. \tag{4.32}$$

The problem for investigating asymptotic behaviors of (4.29) for large time will be left for future work.

D. Dynamical systems related to Eq. (4.1)

In this subsection, we derive the dynamical systems related to Eq. (4.1). The discussion almost parallels that for Eq. (1.1) or that for the ILW equation.[29,30] In so doing, we assume the bilinear variable f defined in (4.3) in the form

$$f = \prod_{j=1}^{M}(x - x_j), \quad |\mathrm{Im}\,x_j| > \delta, \tag{4.33}$$

where the x_j are complex functions of t and M is a positive integer. The conditions $|\mathrm{Im}\,x_j| > \delta$ $(j = 1,2,...,M)$ are required because of the assumption for the analytical property of f [see (4.3) and (4.4)]. Substituting (4.33) into (4.5) and using partial fraction decomposition, we obtain the equation

$$\sum_{j=1}^{M}\left(\frac{1}{x - x_j + i\delta} - \frac{1}{x - x_j - i\delta}\right)(\dot{x}_j - 1) + 2\delta$$

$$\times\sum_{j=1}^{M}\sum_{\substack{l=1\\(l\neq j)}}^{M}\left[\frac{1}{x - x_j + i\delta}\frac{1}{(x_j - x_l)(x_j - x_l - 2i\delta)}\right.$$

$$\left. - \frac{1}{x - x_j - i\delta}\frac{1}{(x_j - x_l)(x_j - x_l + 2i\delta)}\right]$$

$$\times(\dot{x}_j + \dot{x}_l) = 0. \tag{4.34}$$

Then we have, by taking the coefficients of $(x-x_j+i\delta)^{-1}$ and $(x-x_j-i\delta)^{-1}$ zero, respectively, the following system of equations for x_j:

$$\dot{x}_j = 1 - 2\delta \sum_{\substack{l=1\\(l\neq j)}}^{M} \frac{\dot{x}_j+\dot{x}_l}{(x_j-x_l)(x_j-x_l-2i\delta)}, \qquad (4.35\text{a})$$

$$\dot{x}_j = 1 - 2\delta \sum_{\substack{l=1\\(l\neq j)}}^{M} \frac{\dot{x}_j+\dot{x}_l}{(x_j-x_l)(x_j-x_l+2i\delta)}$$

$$(j=1,2,...,M). \qquad (4.35\text{b})$$

Adding (4.5a) and (4.5b) yields

$$\dot{x}_j = 1 - 4\delta \sum_{\substack{l=1\\(l\neq j)}}^{M} \frac{\dot{x}_j+\dot{x}_l}{(x_j-x_l)^2+4\delta^2}, \qquad (4.36\text{a})$$

while subtracting (4.35a) from (4.35b) yields

$$\sum_{\substack{l=1\\(l\neq j)}}^{M} \frac{\dot{x}_j+\dot{x}_l}{(x_j-x_l)\{(x_j-x_l)^2+4\delta^2\}} = 0. \qquad (4.36\text{b})$$

Equation (4.36a) is a finite-dimensional dynamical system with a constraint (4.36b) and it is closely related to the motion of the positions of the poles of solutions of Eq. (4.1). The detailed analysis of Eq. (4.36) will not be done here. Instead, we consider two limiting cases of $\delta\rightarrow\infty$ and $\delta\rightarrow 0$.

1. Deep-water limit: $\delta\rightarrow\infty$

In this limit, it is appropriate to introduce the new variables $\tilde{x}_j$ $(j=1,2,...,M)$ by the relations[29,30]

$$\tilde{x}_j = x_j - i\delta, \quad \text{Im}\, x_j > \delta \quad (j=1,2,...,N), \qquad (4.37\text{a})$$

$$\tilde{x}_j = x_j + i\delta, \quad \text{Im}\, x_j < -\delta \quad (j=N+1,N+2,...,M). \qquad (4.37\text{b})$$

Hence $\tilde{x}_j$, for $j=1,2,...,N$, lie in the upper half plane $(\text{Im}\,\tilde{x}_j>0)$ and $\tilde{x}_j$ for $j=N+1,N+2,...,M$ lie in the lower half plane $(\text{Im}\,\tilde{x}_j<0)$. Substituting (4.37) into Eq. (4.34), we find, in the limit $\delta\rightarrow\infty$,

$$\sum_{j=1}^{N}\frac{1}{x-\tilde{x}_j}\left[\dot{\tilde{x}}_j - 1 + i\sum_{\substack{l=1\\(l\neq j)}}^{N}\frac{\dot{\tilde{x}}_j+\dot{\tilde{x}}_l}{\tilde{x}_j-\tilde{x}_l}\right.$$
$$\left. - i\sum_{l=N+1}^{M}\frac{\dot{\tilde{x}}_j+\dot{\tilde{x}}_l}{\tilde{x}_j-\tilde{x}_l}\right] - \sum_{j=N+1}^{M}\frac{1}{x-\tilde{x}_j}$$
$$\times\left[\dot{\tilde{x}}_j - 1 + i\sum_{l=1}^{N}\frac{\dot{\tilde{x}}_j+\dot{\tilde{x}}_l}{\tilde{x}_j-\tilde{x}_l} - i\sum_{\substack{l=N+1\\(l\neq j)}}^{M}\frac{\dot{\tilde{x}}_j+\dot{\tilde{x}}_l}{\tilde{x}_j-\tilde{x}_l}\right] = 0, \qquad (4.38)$$

whereupon we readily obtain by taking the coefficient of $(x-\tilde{x}_j)^{-1}$ for $j=1,2,...,N$ and that for $j=N+1$, $N+2,...,M$ zero, respectively,

$$\dot{\tilde{x}}_j = 1 - i\sum_{\substack{l=1\\(l\neq j)}}^{N}\frac{\dot{\tilde{x}}_j+\dot{\tilde{x}}_l}{\tilde{x}_j-\tilde{x}_l} + i\sum_{l=N+1}^{M}\frac{\dot{\tilde{x}}_j+\dot{\tilde{x}}_l}{\tilde{x}_j-\tilde{x}_l}$$

$$(j=1,2,...,N), \qquad (4.39\text{a})$$

$$\dot{\tilde{x}}_j = 1 - i\sum_{l=1}^{N}\frac{\dot{\tilde{x}}_j+\dot{\tilde{x}}_l}{\tilde{x}_j-\tilde{x}_l} + i\sum_{\substack{l=N+1\\(l\neq j)}}^{M}\frac{\dot{\tilde{x}}_j+\dot{\tilde{x}}_l}{\tilde{x}_j-\tilde{x}_l}$$

$$(j=N+1,N+2,...,M). \qquad (4.39\text{b})$$

This system of equations is a dynamical system without any constraint. If we take $M=2N$ and $\tilde{x}_{N+j}=\tilde{x}_j^*$ $(j=1,2,...,N)$, Eqs. (4.39a) become

$$\dot{\tilde{x}}_j = 1 - i\sum_{\substack{l=1\\(l\neq j)}}^{N}\frac{\dot{\tilde{x}}_j+\dot{\tilde{x}}_l}{\tilde{x}_j-\tilde{x}_l} + i\sum_{l=1}^{N}\frac{\dot{\tilde{x}}_j+\dot{\tilde{x}}_l^*}{\tilde{x}_j-\tilde{x}_l^*} \quad (j=1,2,...,N), \qquad (4.40)$$

and Eqs. (4.39b) become the complex conjugate expressions of (4.40). Furthermore, in the limit $\delta\rightarrow\infty$, we have from (4.3) and (4.37)

$$u = i\frac{\partial}{\partial x}\ln\left(\frac{\tilde{f}^*}{\tilde{f}}\right), \qquad (4.41\text{a})$$

with

$$\tilde{f} = \sum_{j=1}^{N}(x-\tilde{x}_j), \quad \text{Im}\,\tilde{x}_j>0 \quad (j=1,2,...,N). \qquad (4.41\text{b})$$

The system of equations (4.40) is identical to Eqs. (3.8) with the variables $\tilde{x}_j$ in place of x_j which have already been reduced from the periodic dynamical system related to Eq. (1.1).

2. Shallow-water limit: $\delta\rightarrow 0$

In this limit, it is convenient to introduce the variables $\tilde{t}$ and $\tilde{x}$ defined in (4.10) and the new variables $\tilde{x}_j=(3/\delta)^{1/2}x_j$ $(j=1,2,...,M)$. We then immediately find from (4.36)

$$\dot{\tilde{x}}_j = 1 - 12\sum_{\substack{l=1\\(l\neq j)}}^{M}\frac{\dot{\tilde{x}}_j+\dot{\tilde{x}}_l}{(\tilde{x}_j-\tilde{x}_l)^2}, \qquad (4.42\text{a})$$

$$\sum_{\substack{l=1\\(l\neq j)}}^{M}\frac{\dot{\tilde{x}}_j+\dot{\tilde{x}}_l}{(\tilde{x}_j-\tilde{x}_l)^3} = 0, \qquad (4.42\text{b})$$

which is a dynamical system with a constraint. The dependent variable u, (4.3) now takes the form

$$u = 6\frac{\partial^2}{\partial\tilde{x}^2}\ln\tilde{f}, \qquad (4.43\text{a})$$

with

$$\tilde{f} = \sum_{j=1}^{M}(\tilde{x}-\tilde{x}_j). \qquad (4.43\text{b})$$

The system of equations (4.42) represents a dynamical system related to Eq. (1.2) and the solutions may be constructed from the N-soliton solution (4.19) of Eq. (1.2) by taking an appropriate limiting procedure. The solutions $\tilde{x}_j$ thus obtained are expected to have a form of algebraic functions of $\tilde{t}$ and $\tilde{x}$. However, solutions u themselves constructed from $\tilde{x}_j$ would become singular as seen from a rational solution (4.27), for example, which has been reduced from a one-soliton solution of Eq. (4.1). The situation mentioned here would be the same as that for rational solutions of the KdV equation.[20] The detailed analysis of the system of equations (4.42) will be dealt with elsewhere.

V. CONCLUSION

In this paper, we have proposed a new integrable NIDE that exhibits the N-periodic and N-soliton solutions and showed that it is closely related to certain solvable dynamical systems. The NIDE may be relevant as a model equation that describes wave propagations in fluids of great depth. Moreover, we have generalized our NIDE to a more general one that is an intermediate version between our equation and that of Hirota and Satsuma.[14] The generalized equation may also describe suitably wave phenomena in fluids of finite depth.

In the context of the soliton theory, it is quite interesting to derive the inverse scattering transforms (IST's), the Bäcklund transformations, and an infinite number of conservation laws, etc., for these NIDE's. In this respect, the bilinear equations (2.3) and (4.5) may offer a proper starting point for analyzing these problems since the systematic method for constructing IST's, etc., on the basis of the bilinear equations has already been established considerably within the framework of the bilinear formalism.[16]

Finally, it is useful to comment on another type of integrable model equations for shallow water waves proposed by Ablowitz *et al.*[31] It may read in the form

$$u_t - u_{txx} - 2uu_t + u_x \int_x^\infty u_t\,dx + u_x = 0. \qquad (5.1)$$

Although only the coefficient of the nonlinear term uu_t of Eq. (5.1) differs in comparison with Eq. (1.2), Eq. (5.1) is never reducible to Eq. (1.2) by means of any scale transformations. If we replace the dispersive term u_{txx} by Hu_{tx}, new NIDE's will arise. The method for exact solution developed in this paper may be applied to these NIDE's in order to obtain various results corresponding to those presented here. A number of problems proposed in this paper will be dealt with in the near future.

ACKNOWLEDGMENT

The author would like to thank Professor M. Nishioka for his kind encouragement.

[1] Y. Matsuno, J. Phys. A: Math. Gen. **12**, 619 (1979); **13**, 1519 (1980).

[2] J. Satsuma and Y. Ishimori, J. Phys. Soc. Jpn. **46**, 681 (1979).

[3] K. M. Case, Proc. Natl. Acad. Sci. USA **75**, 3562 (1978); **76**, 1 (1979).

[4] H. H. Chen, Y. C. Lee, and N. R. Pereira, Phys. Fluids **22**, 187 (1979).

[5] A. S. Fokas and M. J. Ablowitz, Stud. Appl. Math. **68**, 1 (1983).

[6] Y. Matsuno, Phys. Lett. A **74**, 223 (1979).

[7] H. H. Chen and Y. C. Lee, Phys. Rev. Lett. **43**, 264 (1979).

[8] Y. Kodama, M. J. Ablowitz, and J. Satsuma, J. Math. Phys. **23**, 564 (1982).

[9] A. Degasperis and P. M. Santini, Phys. Lett. A **98**, 240 (1983).

[10] A. Degasperis, P. M. Santini, and M. J. Ablowitz, J. Math. Phys. **26**, 2469 (1985).

[11] Y. Matsuno, Phys. Lett. A **119**, 229 (1986); **120**, 187 (1987); J. Phys. A: Math. Gen. **20**, 3587 (1987).

[12] P. Constantin, P. D. Lax, and A. Majada, Commun. Pure Appl. Math. **38**, 715 (1985).

[13] M. J. Ablowitz, A. S. Fokas, and M. D. Kruskal, Phys. Lett. A **120**, 215 (1987).

[14] R. Hirota and J. Satsuma, J. Phys. Soc. Jpn. **40**, 611 (1976).

[15] R. Hirota, Phys. Rev. Lett. **27**, 1192 (1971).

[16] Y. Matsuno, *Bilinear Transformation Method* (Academic, New York, 1984).

[17] T. B. Benjamin, J. Fluid Mech. **29**, 559 (1967).

[18] H. Ono, J. Phys. Soc. Jpn. **39**, 1082 (1975).

[19] M. D. Kruskal, Lect. Appl. Math., Am. Math. Soc. **15**, 61 (1974).

[20] H. Airault, H. P. Mckean, and J. Moser, Commun. Pure Appl. Math. **30**, 95 (1977).

[21] D. V. Choodnovsky and G. V. Choodnovsky, Nuovo Cimento B **40**, 339 (1977).

[22] F. Calogero, Nuovo Cimento B **43**, 177 (1978).

[23] M. A. Olshanetsky and A. M. Perelomov, Phys. Rep. **71**, 313 (1981).

[24] R. I. Joseph, J. Phys. A: Math. Gen. **10**, L225 (1977).

[25] R. I. Joseph and R. Egri, J. Phys. A: Math. Gen. **11**, L97 (1978).

[26] T. Kubota, D. R. S. Ko, and D. Dobbs, J. Hydronaut. **23**, 157 (1978).

[27] A. Nakamura and Y. Matsuno, J. Phys. Soc. Jpn. **48**, 653 (1980).

[28] M. J. Ablowitz and J. Satsuma, J. Math. Phys. **19**, 2180 (1978).

[29] J. Satsuma and M. J. Ablowitz, *Nonlinear Partial Differential Equations in Engineering and Applied Science*, edited by R. L. Sternberg, A. J. Kalinowski, and J. S. Papadakis (Dekker, New York, 1980).

[30] M. J. Ablowitz and H. Segur, *Solitons and the Inverse Scattering Transform* (SIAM, Philadelphia, 1981).

[31] M. J. Ablowitz, D. J. Kaup, A. C. Newell, and H. Segur, Stud. Appl. Math. **53**, 249 (1974).

Special relativity with Clifford algebras and 2×2 matrices, and the exact product of two boosts

W. E. Baylis and George Jones
Department of Physics, University of Windsor, Windsor, Ontario, N9B 3P4, Canada

(Received 26 August 1986; accepted for publication 19 August 1987)

Formulations of the special theory of relativity in the Dirac or "space-time" algebra are compared with those in the simpler algebra of 2×2 matrices ("Pauli algebra"). The Dirac algebra separates elements into odd and even multivectors, but this feature is not needed in most practical calculations. As a result, Pauli-algebra formulations are just as powerful in most cases. Furthermore, the new correction angle ϕ, which Salingaros found with the Dirac algebra to be required to describe the product of two boosts, is shown to be identically zero, and new results for special boost combinations are derived.

I. INTRODUCTION

Clifford algebras have become popular tools in the description of relativistic physics.[1-7] Such algebras deal with multivectors which can be formed from linear combinations of antisymmetric products of a set of basis vectors. Salingaros,[3-7] a prolific proponent of Clifford algebras, has pointed out their advantages, especially the facility they allow in making coordinate-free calculations of finite Lorentz transformations. The space-time or Dirac algebra[1,2] (see Sec. III, below), constructed from a set of four basis vectors on Minkowski space, seems the natural choice for applications in relativistic physics. However, as we discuss below, the Dirac algebra, with its basis of 16 independent forms over the real numbers and with multiplication rules unfamiliar to most physicists, is unnecessarily complicated for most calculations in special relativity. It can usually be replaced by the Pauli algebra, which has a simple representation as the algebra of 2×2 matrices.[8]

In a recent paper, Salingaros[5] used the Clifford algebra of space-time (Dirac algebra) to calculate the product of two boost transformations and to express the result as a product of a net boost and a rotation. He claimed that his Clifford-algebra techniques allowed him to derive an exact result which differs by a rotation of the net boost direction from the standard result as given, say, by Møller[9] or Jackson.[10] The standard result is usually obtained by calculating velocities from coordinates related by a boost transformation, or equivalently, by boosting a four-velocity.[8]

In an erratum Salingaros[5] corrected an algebraic error but still maintained that the physical predictions of the Pauli and Dirac algebras are in disagreement. We show below that the predictions of both algebras are identical for the product of boosts, and that consequently, it is not necessary to address the question of "which Clifford algebra correctly describes the physical Lorentz group." [5]

The remainder of this paper is in three parts. First, in Sec. II, we investigate constraints on the products of boosts that arise from general symmetry arguments. Next, in Sec. III, we relate the familiar algebra of 2×2 matrices to the Clifford algebra used by Salingaros[3-7] and others[1] and show why, for practical calculations in special relativity, the former is nearly of equal power in spite of its much greater simplicity. In Sec. IV, we use the simpler algebra to demonstrate the identity of the net boost velocity from the product of two boosts to that found by directly boosting a four-velocity. This procedure is then shown to give results in complete accord with those derived from the Dirac algebra. We also derive the exact expression for the net boost when sandwiched between identical half-angle rotations [corresponding to s in Eq. (40) of Ref. 5] in order to demonstrate the symmetry properties predicted in Sec. II, and we show how boosts may be combined to produce net boosts with no rotation. The "inverse problem" for pure boosts, namely to find the boost in terms of given initial and transformed four-vectors or "six-vectors," is solved as a further example of the power of Pauli-algebra techniques in relativistic kinematics. In the Conclusions (Sec. V), we summarize some of the evidence concerning the relative usefulness of the two Clifford algebras for applications in special relativity.

II. SOME SYMMETRY PROPERTIES OF PRODUCTS OF LORENTZ TRANSFORMATIONS

The group $L^{\uparrow}_{+}$ of restricted Lorentz transformations includes both pure boosts $B(\mathbf{w})$ and pure rotations $R(\boldsymbol{\Omega})$, where the boost parameter $\mathbf{w}$ is simply related to the relative velocity

$$\mathbf{v} = \mathbf{w}(\tanh w)/w, \quad w = |\mathbf{w}|, \tag{1}$$

induced by the boost and the vector angle $\boldsymbol{\Omega}$ gives both the magnitude $\Omega = |\boldsymbol{\Omega}|$ of the angle of rotation and the direction $\hat{\Omega} = \boldsymbol{\Omega}/\Omega$ of the right-handed axis of rotation. The rotations form a subgroup SO(3) of $L^{\uparrow}_{+}$, but the boosts are not a group in themselves because the product of two boosts, say, $B(\mathbf{w}_1)$ followed by $B(\mathbf{w}_2)$, is equivalent to a net boost $B(\mathbf{w})$ followed by a rotation $R(\boldsymbol{\Omega})$:

$$B(\mathbf{w}_2)B(\mathbf{w}_1) = R(\boldsymbol{\Omega})B(\mathbf{w}). \tag{2}$$

In this paper, unless otherwise specified, we mean *active* transformations, i.e., of physical systems, rather than *passive* ones, of coordinate frames. The two are related by changes of sign in the parameters, but the distinction is not significant for most of the arguments of this section.

Other decompositions are possible since it follows from Eq. (2) that

0022-2488/88/010057-06$02.50

$$
\begin{aligned}
B(\mathbf{w}_2)B(\mathbf{w}_1) &= [R(\boldsymbol{\Omega})B(\mathbf{w})R^{-1}(\boldsymbol{\Omega})]R(\boldsymbol{\Omega}) \\
&= B(\mathbf{w}')R(\boldsymbol{\Omega}) \qquad (2') \\
&= R(\boldsymbol{\Omega}/2)B(\mathbf{w}'')R(\boldsymbol{\Omega}/2), \qquad (2'')
\end{aligned}
$$

where $R^{-1}(\boldsymbol{\Omega}) = R(-\boldsymbol{\Omega})$ is the inverse of $R(\boldsymbol{\Omega})$ and where the parameters $\mathbf{w}'$ and $\mathbf{w}''$ are obtained from $\mathbf{w}$ by rotations of $\boldsymbol{\Omega}$ and $\boldsymbol{\Omega}/2$, respectively. The inverse transformation is

$$
\begin{aligned}
[B(\mathbf{w}_2)B(\mathbf{w}_1)]^{-1} &= B(-\mathbf{w}_1)B(-\mathbf{w}_2) \\
&= B(-\mathbf{w})R(-\boldsymbol{\Omega}). \qquad (3)
\end{aligned}
$$

Spatial inversion changes the sign of the boost parameters but not the direction of the rotation axis, so that Eq. (3) becomes

$$B(\mathbf{w}_1)B(\mathbf{w}_2) = B(\mathbf{w})R(-\boldsymbol{\Omega}). \qquad (4)$$

A comparison of Eqs. (4) and (2′) shows immediately that once $\mathbf{w}$ has been determined as an analytic function $\mathbf{w}(\mathbf{w}_1,\mathbf{w}_2)$ of $\mathbf{w}_1$ and $\mathbf{w}_2$, $\mathbf{w}'$ is given by simply interchanging the indices 1 and 2. Furthermore, the sign of $\boldsymbol{\Omega} = \boldsymbol{\Omega}(\mathbf{w}_1,\mathbf{w}_2)$ should change when the indices indicating the order of application of the component boosts are switched:

$$\mathbf{w}'(\mathbf{w}_1,\mathbf{w}_2) = \mathbf{w}(\mathbf{w}_2,\mathbf{w}_1), \qquad (5a)$$

$$\boldsymbol{\Omega}(\mathbf{w}_1,\mathbf{w}_2) = -\boldsymbol{\Omega}(\mathbf{w}_2,\mathbf{w}_1). \qquad (5b)$$

Similarly, the comparison of Eq. (2″) with the inverse of its spatial inversion,

$$B(\mathbf{w}_1)B(\mathbf{w}_2) = R(-\boldsymbol{\Omega}/2)B(\mathbf{w}'')R(-\boldsymbol{\Omega}/2), \qquad (6)$$

demonstrates, in addition to Eq. (5b), that $\mathbf{w}'' = \mathbf{w}''(\mathbf{w}_1,\mathbf{w}_2)$ is symmetric with respect to the interchange $\mathbf{w}' \leftrightarrow \mathbf{w}_2$,

$$\mathbf{w}''(\mathbf{w}_1,\mathbf{w}_2) = \mathbf{w}''(\mathbf{w}_2,\mathbf{w}_1). \qquad (7)$$

This constraint, for example, proves that Eq. (2″) cannot be satisfied when $\mathbf{w}''$ is replaced by the standard result $\mathbf{s}$ for the net boost [Eqs. (33) and (34) and (40) of Ref. (5) (before correction in the Erratum)], because $\mathbf{s}$ is clearly not symmetric with respect to the order in which the boosts are applied. In Sec. IV we derive exact results for $\mathbf{w}''$ and show that the symmetry condition equation (7) is indeed fulfilled.

Another interesting way to combine boosts is to boost by $\mathbf{w}_1/2$ both before and after a boost by $\mathbf{w}_2$. Assuming the result to be equivalent to a boost by $\mathbf{w}_3$ preceded and followed by a unique rotation of $\boldsymbol{\Omega}_3/2$, we write, in analogy to Eq. (2′),

$$B(\mathbf{w}_1/2)B(\mathbf{w}_2)B(\mathbf{w}_1/2) = R(\boldsymbol{\Omega}_3/2)B(\mathbf{w}_3)R(\boldsymbol{\Omega}_3/2). \qquad (8)$$

A comparison of Eq. (8) with the transformation inverse of its spatial inverse gives

$$
\begin{aligned}
&R(\boldsymbol{\Omega}_3/2)B(\boldsymbol{\omega}_3)R(\boldsymbol{\Omega}_3/2) \\
&\quad = R(-\boldsymbol{\Omega}_3/2)B(\mathbf{w}_3)R(-\boldsymbol{\Omega}_3/2). \qquad (9)
\end{aligned}
$$

Consequently the rotation angle $\boldsymbol{\Omega}_3$ must be zero and the combination equation (8) must be a pure boost. The result is also easily derived from a different approach within the algebra of 2×2 matrices, as is shown in Sec. IV.

III. DIRAC ALGEBRA VERSUS THE ALGEBRA OF 2×2 MATRICES

In Refs. 4–6 Salingaros used the Clifford algebra of multivectors formed from products of vectors in four-dimensional Minkowski space. The space-time algebra of Hestenes[1] is equivalent. In this section we compare this algebra with the simpler algebra of 2×2 matrices for calculations in special relativity.[8] Of course, one's choice of the "best" algebra is largely a subjective matter of personal preference, but a number of relevant facts can be ascertained. First, we must briefly describe the algebras we want to compare. Only the essentials needed to clarify our notation and comparisons are included here. Further details about the Clifford algebra may be found in Refs. 1–7.

From the four basis vectors, say $\mathbf{e}_\mu, \mu = 0,1,2,3$, of Minkowski space, one builds a basis of 16 independent forms from all possible antisymmetric outer products $\mathbf{e}_\lambda \wedge \mathbf{e}_\mu \cdots$. The 4×4 Dirac matrices γ^μ and their products provide a representation of the basis forms, and indeed this Clifford algebra is often referred to as the Dirac algebra[1,2] $\mathscr{D}$. (Some authors reserve the name Dirac algebra for the algebra with 32 independent basis forms resulting from the complexification of our $\mathscr{D}$.[7]) Any element of $\mathscr{D}$ is thus a linear combination with real coefficients of a scalar, the four basis vectors $\mathbf{e}_\mu$, the six basis bivectors $\mathbf{e}_\mu \wedge \mathbf{e}_\nu$, the four pseudovectors $\mathbf{e}_\lambda \wedge \mathbf{e}_\mu \wedge \mathbf{e}_\nu$, and the pseudoscalar $\omega = \mathbf{e}_0 \wedge \mathbf{e}_1 \wedge \mathbf{e}_2 \wedge \mathbf{e}_3$, $\omega^2 = -1$.

The scalar, bivector, and pseudoscalar parts of an element of $\mathscr{D}$ are said to be "even," whereas the vector and pseudovector parts are "odd." The bilinear products of two even or two odd elements is even whereas the product of an even with an odd element is odd. The even elements of $\mathscr{D}$ form a subalgebra $\mathscr{D}_+$ which is isomorphic to the Pauli algebra $\mathscr{P}$, the Clifford algebra of products of vectors in three-dimensional Euclidean space.[1–3] The subspace $\mathscr{D}_-$ of odd elements of $\mathscr{D}$ is not an algebra.

If the Euclidean basis vectors are written $\boldsymbol{\sigma}_k$, $k = 1,2,3$, then the eight basis forms of $\mathscr{P}$ are $\{1, \boldsymbol{\sigma}_k, i\boldsymbol{\sigma}_k, i\}$. The name Pauli algebra originates from the simple representation of $\boldsymbol{\sigma}_k$ by the Pauli spin matrices

$$\underline{\sigma}_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \underline{\sigma}_2 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \quad \underline{\sigma}_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}. \qquad (10)$$

Since $\underline{1} = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$ and $\underline{\sigma}_k$, $k = 1,2,3$, form a complete linearly independent set of 2×2 matrices, $\mathscr{P}$ is seen to be isomorphic to the algebra of 2×2 matrices with complex elements. The latter is undeniably simpler—but also more restricted—than $\mathscr{D}$ because the number of basis forms is reduced by half. It is also more familiar to most physicists because the ring multiplication is given by the usual multiplication of 2×2 matrices, and inner and outer products can be replaced by the usual "dot" and "cross" products of vectors. For example, the product of two complex vectors of the form $\underline{a} = a^k\underline{\sigma}_k$ (the summation convention is assumed for repeated indices and underscores indicate 2×2 matrices) gives

$$\underline{\mathbf{ab}} = \underline{\mathbf{a}}\cdot\underline{\mathbf{b}} + i\underline{\mathbf{a}}\times\underline{\mathbf{b}} = \mathbf{a}\cdot\mathbf{b}\underline{1} + i(\mathbf{a}\times\mathbf{b})^k\underline{\sigma}_k, \qquad (11)$$

where $\mathbf{a}\cdot\mathbf{b} = a^k b^k$ and $(\mathbf{a}\times\mathbf{b})^k = \epsilon_{ijk}a^i b^j$ are the normal dot and cross products.

Since the Dirac algebra comprises products of Minkowski-space vectors, it at first appears the one best adapted for work in special relativity. However, it can easily be shown to be more general than usually needed. Elements of $\mathscr{D}$ can be separable mixtures of even and odd multivectors, but one rarely needs such mixtures in practical calculations. The reason is that physical quantities correspond to either purely even or purely odd elements of $\mathscr{D}$, and multiplication of such elements together always yields again purely even or purely odd elements.[11] There is no need to accommodate mixtures. By mapping even and odd parts onto the same subalgebra, the number of basis forms can be halved.

As discussed above, the Pauli algebra and hence the algebra of 2×2 complex matrices is isomorphic to the subalgebra $\mathscr{D}_+$ of even elements of $\mathscr{D}$. The vectors and pseudovectors of $\mathscr{D}_-$ can be mapped onto $\mathscr{D}_+$ by multiplying them by $\mathbf{e}_0$: the $\mathbf{e}_0$ component of each vector becomes a scalar and the $\mathbf{e}_k$ components, $k = 1,2,3$, become bivectors $\mathbf{e}_0 \wedge \mathbf{e}_k$, whereas the $\mathbf{e}_1 \wedge \mathbf{e}_2 \wedge \mathbf{e}_3$ component of each pseudovector becomes a pseudoscalar and the other components become bivectors $\mathbf{e}_j \wedge \mathbf{e}_k$. Thus both odd and even parts of $\mathscr{D}$ can be mapped onto $\mathscr{D}_+$, and hence onto $\mathscr{P}$ and onto the algebra of 2×2 matrices. Since the mapping is one-to-one for both the even part and the odd part of $\mathscr{D}$, the much simpler algebra of 2×2 matrices has all the power of $\mathscr{D}$ in calculations which do not require mixed even/odd elements.

The advantages of $\mathscr{P}$ (we will not usually distinguish between $\mathscr{P}$ and its representation, the algebra of 2×2 matrices) in comparison to the Dirac algebra stem mainly from its relative simplicity, its explicit representation, and the familiarity of physicists with the algebraic operations. One can avoid the inner and outer products of $\mathscr{D}$, whose symmetry relationships to the basic ring multiplication (Salingaros's "vee" product[5,6]) depend on the element types being multiplied. One also avoids a pseudoscalar which anticommutes with vectors and pseudovectors. One gains a more direct separation of four-space vectors into time and space components, and in three-space, vectors are easily distinguished from pseudovectors, which here (but not in $\mathscr{D}$) share their familiar identity with axial vectors.

The algebra of 2×2 matrices is also isomorphic to the algebra of complex quaternions, and the advantages of applying such an algebra to problems in special relativity (as well as for pure rotations) were discussed in print as early as 1910 by Klein and Sommerfeld.[12] Many authors[8,13-16] have contributed to the theory, including Salingaros and Ilamed.[17] When 2×2 matrices are used to represent "four-vectors" of Minkowski space-time, say,

$$\underline{x} = x_0\underline{1} + \underline{\mathbf{x}} = x_0\underline{1} + x^k\underline{\sigma}_k, \tag{12}$$

Lorentz transformations

$$\underline{x} \rightarrow \underline{x}' = \underline{L}\,\underline{x}\,\underline{L}^{+} \tag{13}$$

are determined directly by unimodular 2×2 matrices of SL(2,C) which can generally be written[4]

$$\underline{L} = \exp[(\underline{\mathbf{w}} - i\underline{\mathbf{\Omega}})/2]. \tag{14}$$

If $\mathbf{w} = 0$, the transformation is a pure rotation and the matrix $\underline{L} = \underline{R}$ is unitary, whereas if $\mathbf{\Omega} = 0$, the transformation is a pure boost and $\underline{L} = \underline{B}$ is Hermitian:

$$\begin{aligned} \underline{R}(\mathbf{\Omega}) &= \underline{R}^{+}(-\mathbf{\Omega}) = \exp(-i\underline{\mathbf{\Omega}}/2), \\ \underline{B}(\mathbf{w}) &= \underline{B}^{+}(\mathbf{w}) = \exp(\underline{\mathbf{w}}/2). \end{aligned} \tag{15}$$

If $\underline{x}$ and $\underline{p} = p_0\underline{1} + \underline{\mathbf{p}}$ are any two four-vectors, the product of one with the spatial reversal,[18] say,

$$\bar{\underline{p}} = p_0\underline{1} - \underline{\mathbf{p}}, \tag{16}$$

of the other gives matrices with still simple but distinct transformation rules:

$$\underline{x}\bar{\underline{p}} \rightarrow \underline{L}\,\underline{x}\,\underline{L}^{+}\underline{L}^{+}\bar{\underline{p}}\overline{\underline{L}} = \underline{L}\,\underline{x}\bar{\underline{p}}\overline{\underline{L}}, \tag{17}$$

where we have noted that the spatial reversal of $\underline{L}$ is also its matrix inverse,

$$\overline{\underline{L}} = \underline{L}^{-1}. \tag{18}$$

Ones sees that the scalar part of $\underline{x}\,\bar{\underline{p}}$ is Lorentz invariant (it corresponds to the usual scalar product of four-vectors) and that the complex-vector (six-vector) part of $\underline{x}\,\bar{\underline{p}}$, although it transforms under rotation just like the vector part of a four-vector, transforms distinctly under boosts. Angular momentum, electric and magnetic fields, and the transformation parameters $\mathbf{w}$ and $\mathbf{\Omega}$ themselves are parts of six-vectors, and the 2×2 matrix formalism easily explains why they transform differently under boosts than do four-vectors. Clearly higher-order products of four-vectors can be formed which transform either like four-vectors (odd multivectors) or like a scalar plus a six-vector (even multivectors). Further details and applications of 2×2-matrix algebra to special relativity are given in Ref. 8.

Lorentz transformations in the Dirac algebra $\mathscr{D}$ are quite similar in structure to those in 2×2-matrix algebra $\mathscr{P}$, but in $\mathscr{D}$ the form is exactly the same for all elements. Recall, however, that the multiplication rules for inner and outer products change when a vector is replaced, say, by a bivector, so that even though the form appears the same, the results of Lorentz transformations can differ.

The bivectors of $\mathscr{D}$ correspond to the six-vectors of $\mathscr{P}$, but in $\mathscr{D}$ the bivectors and vectors are clearly distinguishable because they span nonoverlapping subsets of basis forms whereas in $\mathscr{P}$, because of the mapping of both $\mathscr{D}_+$ and $\mathscr{D}_-$ onto $\mathscr{P}$, they span overlapping subsets. Consequently in $\mathscr{P}$, one must know whether a given spatial vector is part of a four-vector (such as a vector potential) or part of a six-vector (such as an electric field) in order to know its transformation properties under boosts. However, this is a familiar and rather trivial problem, since, as pointed out above, there is no need in practice to superimpose even and odd multivectors.

IV. NET BOOSTS

The algebra $\mathscr{P}$ of 2×2 matrices can be demonstrated by deriving several relevant results. First, we show the relationship between a boosted four-velocity and a net boost.

The four-velocity $\underline{u}$ in the rest frame has only a timelike component. In units with the velocity of light $c = 1$, it is

$$\underline{u}_0 = \underline{1}. \tag{19}$$

If the system at rest is boosted by $B(\mathbf{w})$ [see Eqs. (13)–(15)], its new four-velocity is

$$\underline{u} = \underline{B}(\mathbf{w})\underline{1}\,\underline{B}^{+}(\mathbf{w}) = e^{\underline{\mathbf{w}}}. \tag{20}$$

Thus one sees that the four-velocity is identical to the square of the 2×2 matrix representing the corresponding boost. By power-series expansion of $e^{\underline{\mathbf{w}}}$ and applications of Eq. (11), one obtains

$$\underline{u} = \underline{1}\cosh w + \underline{\mathbf{w}}(\sinh w)/w, \tag{21}$$

where $w = (\mathbf{w}\cdot\mathbf{w})^{1/2}$. Identifying

$$\gamma = \cosh w, \quad \mathbf{u} = \gamma\mathbf{v} = \mathbf{w}(\sinh w)/w, \tag{22}$$

we write

$$\underline{u} = \gamma\underline{1} + \underline{\mathbf{u}}, \quad \gamma = (1+u^2)^{1/2} = (1-v^2)^{-1/2}. \tag{23}$$

Now take a system with four-velocity $\underline{u}_1 = \underline{B}(\mathbf{w}_1)\underline{B}^+(\mathbf{w}_1)$ and boost it by $\underline{B}(\mathbf{w}_2)$:

$$\begin{aligned}\underline{u}_1 \to \underline{u}' &= \underline{B}(\mathbf{w}_2)\underline{u}_1\underline{B}^+(\mathbf{w}_2)\\ &= \underline{B}(\mathbf{w}_2)\underline{B}(\mathbf{w}_1)\underline{B}^+(\mathbf{w}_1)\underline{B}^+(\mathbf{w}_2)\\ &= \underline{B}(\mathbf{w}')\underline{B}^+(\mathbf{w}'),\end{aligned} \tag{24}$$

where the net boost $\underline{B}(\mathbf{w}')$ is related to boosts $\underline{B}(\mathbf{w}_1)$ and $\underline{B}(\mathbf{w}_2)$ as in Eq. (2'),

$$\underline{B}(\mathbf{w}_2)\underline{B}(\mathbf{w}_1) = \underline{B}(\mathbf{w}')\underline{R}(\mathbf{\Omega}). \tag{2'}$$

Consequently, results for the net boost $\mathbf{w}'$ or the corresponding net four-velocity $\underline{u}' = \exp\underline{\mathbf{w}}'$ should be identical whether one boosts a four-velocity or directly multiplies two boost transformations. In particular, the spatial directions of $\mathbf{u}'$ and $\mathbf{w}'$ should be the same.

The calculation is trivial in $\mathscr{P}$.[8] One notes from Eq. (11) that three-vectors commute if parallel, and anticommute if perpendicular. Therefore, splitting $\mathbf{u}_1$ into parts parallel and perpendicular to $\mathbf{w}_2$,

$$\mathbf{u}_1 = \mathbf{u}_{1\parallel} + \mathbf{u}_{1\perp}, \quad \mathbf{u}_{1\parallel} = \mathbf{u}_1\cdot\hat{w}_2\hat{w}_2, \tag{25}$$

one finds

$$\begin{aligned}\underline{u}' = e^{\underline{\mathbf{w}}_2/2}\underline{u}_1 e^{\underline{\mathbf{w}}_2/2} &= e^{\underline{\mathbf{w}}_2}(\gamma_1\underline{1} + \underline{\mathbf{u}}_{1\parallel}) + \underline{\mathbf{u}}_{1\perp}\\ &= (\gamma_2\underline{1} + \underline{\mathbf{u}}_2)(\gamma_1\underline{1} + \underline{\mathbf{u}}_{1\parallel}) + \underline{\mathbf{u}}_{1\perp}\\ &= (\gamma_1\gamma_2 + \mathbf{u}_1\cdot\mathbf{u}_2)\underline{1} + (\gamma_1\underline{\mathbf{u}}_2 + \gamma_2\underline{\mathbf{u}}_{1\parallel}) + \underline{\mathbf{u}}_{1\perp}.\end{aligned} \tag{26}$$

The standard results follow immediately when one equates coefficients of the basis matrices $\underline{1}$ and $\underline{\sigma}_k$,

$$\gamma' = \gamma_1\gamma_2 + \mathbf{u}_1\cdot\mathbf{u}_2 = \gamma_1\gamma_2(1 + \mathbf{v}_1\cdot\mathbf{v}_2), \tag{27a}$$

$$\mathbf{v}_1 = \frac{\mathbf{u}'}{\gamma} = \frac{\mathbf{v}_1 + \mathbf{v}_2 + (1-\gamma_2^{-1})\hat{v}_2\times(\hat{v}_2\times\mathbf{v}_1)}{1 + \mathbf{v}_1\cdot\mathbf{v}_2}. \tag{27b}$$

On the other hand, for the net boost which precedes the rotation as in Eq. (2),

$$\begin{aligned}\underline{u} = e^{\underline{\mathbf{w}}} &= \underline{B}^+(\mathbf{w})\underline{R}^+(\mathbf{\Omega})\underline{R}(\mathbf{\Omega})\underline{B}(\mathbf{w})\\ &= \underline{B}^+(\mathbf{w}_1)\underline{B}^+(\mathbf{w}_2)\underline{B}(\mathbf{w}_2)\underline{B}(\mathbf{w}_1)\\ &= \underline{B}(\mathbf{w}_1)\underline{u}_2\underline{B}(\mathbf{w}_1),\end{aligned} \tag{28}$$

so that $\underline{u}$, $\mathbf{w}$, $\mathbf{v}$, $\mathbf{u}$, and γ can be obtained from $\underline{u}'$, $\mathbf{w}'$, $\mathbf{v}'$, $\mathbf{u}'$, and γ', respectively, simply by switching the indices 1 and 2 as predicted by Eq. (5a).

The method of boosting a four-velocity to find the net boost or velocity does not give the rotation angle $\mathbf{\Omega}$ of Eq. (2) or (2') but it does give correctly the magnitude and direction of $\mathbf{w}$ and $\mathbf{w}'$. The angle can be found directly by solving Eq. (2) with the boost and rotation matrices of Eq. (15),

$$e^{\underline{\mathbf{w}}_2/2}e^{\underline{\mathbf{w}}_1/2} = e^{-i\underline{\Omega}/2}e^{\underline{\mathbf{w}}/2}. \tag{29}$$

From the exponential expansions

$$\begin{aligned}e^{\underline{\mathbf{w}}/2} &= \underline{1}\cosh w/2 + \underline{\mathbf{w}}\sinh w/2\\ &= [2(\gamma+1)]^{-1/2}[(\gamma+1)\underline{1} + \underline{\mathbf{u}}],\end{aligned} \tag{30}$$

$$e^{-i\underline{\Omega}/2} = 1\cos\Omega/2 - i\hat{\underline{\Omega}}\sin\Omega/2,$$

one obtains four equations, equivalent to the equality of Eqs. (12) and (13) of Ref. 1, which can be solved for $\mathbf{w}$ and $\mathbf{\Omega}$. The results, as shown, for example, by McFarlane[11] and van Wyk,[12] give $\mathbf{w}$ fully consistent with Eq. (27) above, and yield $\mathbf{\Omega}$ given by

$$\begin{aligned}\cos\frac{\Omega}{2} &= \frac{1+\gamma_1+\gamma_2+\gamma}{[2(\gamma_1+1)(\gamma_2+1)(\gamma+1)]^{1/2}},\\ \hat{\Omega}\sin\frac{\Omega}{2} &= \frac{\mathbf{u}_1\times\mathbf{u}_2}{[2(\gamma_1+1)(\gamma_2+1)(\gamma+1)]^{1/2}}.\end{aligned} \tag{31}$$

These results for $\mathbf{\Omega}$ were given by Hestenes[1] and McFarlane[14] and are consistent with—but simpler than—Eq. (18) of Ref. 5. They are most easily obtained after $\mathbf{w}$ is known by equating scalar parts of and pseudovector parts of

$$\begin{aligned}e^{-i\underline{\Omega}/2} &= e^{\underline{\mathbf{w}}_2/2}e^{\underline{\mathbf{w}}_1/2}e^{-\underline{\mathbf{w}}/2}\\ &= \tfrac{1}{2}[2(\gamma_1+1)(\gamma_2+1)(\gamma+1)]^{-1/2}\\ &\quad\times[(1+\gamma_2)\underline{1} + \underline{\mathbf{u}}_2][(1+\gamma_1)\underline{1} + \underline{\mathbf{u}}_1]\\ &\quad\times[(1+\gamma)\underline{1} - \underline{\mathbf{u}}].\end{aligned} \tag{32}$$

Hestenes[1] [see his Eqs. (18.27) and (18.29) and note a sign error in Eq. (18.27)] derived his results from Dirac-algebra techniques equivalent to those of Salingaros.

Within his erratum[5] Salingaros's results are also in accord with the above analysis. However, he still argues that the Dirac algebra gives physical results that differ from the standard ones found, for example, by means of the Pauli algebra. Indeed, he has derived a correction angle ϕ by which the standard results differ from his own, and he has asserted that generally $\phi \neq 0$. It is therefore important to stress that Clifford-algebra techniques are used in Salingaros's derivation only to establish the four initial equations resulting from a comparison of his Eqs. (12) and (13) and, equivalently, those resulting from his Eq. (26). The rest of the derivation uses only standard algebraic and trigonometric methods to solve for the unknown rotation angle and boost parameter: no techniques of Clifford algebra make any further contribution. Salingaros's four initial equations are identical to equations derived from the Pauli algebra.

The equation given by Salingaros [Eq. (3) in his erratum] does correctly give the additional rotation angle ϕ predicted by the Dirac algebra, but it is not difficult to show, with Eq. (31) above and Salingaros's Eq. (34), that $\phi \equiv 0$ for all combinations of $\mathbf{V}_1$ and $\mathbf{V}_2$, Salingaros's assertions to the contrary notwithstanding. The results of the Pauli and Dirac algebras are thus identical, at least for the product of arbitrary boosts, in conformity with their close relationship as discussed in Sec. III.

The Pauli algebra can be similarly used to derive other expressions in an efficient straightforward way. For example, Eq. (6) becomes

$$e^{\underline{\mathbf{w}}_1/2}e^{\underline{\mathbf{w}}_2/2} = e^{i\underline{\Omega}/2}e^{\underline{\mathbf{w}}''/2}e^{i\underline{\Omega}/2}. \tag{33}$$

Expansions like Eq. (30) give directly

$$\mathbf{u}'' = \mathbf{w}'' \frac{\sinh w}{w} = \left[\frac{\gamma + 1}{2(\gamma_1 + 1)(\gamma_2 + 1)}\right]^{1/2} \times [(\gamma_2 + 1)\mathbf{u}_1 + (\gamma_1 + 1)\mathbf{u}_2], \qquad (34)$$

where $\gamma = \gamma_1\gamma_2 + \mathbf{u}_1\cdot\mathbf{u}_2$. The result (34) does display the symmetry under exchange of indices 1↔2 as predicted in Sec. II. To solve for $\mathbf{u}_3 = \mathbf{w}_3(\sinh w_3)/w_3$ in the case of the rotation-free combination of noncollinear boosts, Eq. (8) may be written

$$e^{\mathbf{w}_3/2} = e^{\mathbf{w}_2/4}e^{\mathbf{w}_1/2}e^{\mathbf{w}_2/4} = e^{\mathbf{w}_2/2}e^{\mathbf{w}_1/2} + (e^{\mathbf{w}_2/2} - 1)[2(\gamma_1 + 1)]^{-1/2}\underline{\hat{u}_2\times(\hat{u}_2\times\mathbf{u}_1)}, \qquad (35)$$

which gives

$$\gamma_3 = \frac{(1 + \gamma_1 + \gamma_2 + \gamma)^2}{2(\gamma_1 + 1)(\gamma_2 + 1)} - 1, \qquad (36)$$

$$\mathbf{u}_3 = \frac{1}{2}(1 + \gamma_1 + \gamma_2 + \gamma)\left\{\frac{\mathbf{u}_1}{\gamma_1 + 1} + \frac{\mathbf{u}_2}{\gamma_2 + 1} + \frac{(\mathbf{u}_1\times\hat{u}_2)\times\hat{u}_2}{\gamma_1 + 1}\left[1 - \left(\frac{2}{\gamma_2 + 1}\right)^{1/2}\right]\right\}. \qquad (37)$$

The fact that the combination $\underline{B}(\mathbf{w}_2/2)\underline{B}(\mathbf{w}_1)\underline{B}(\mathbf{w}_2/2)$ is a pure boost follows in the Pauli algebra simply from its Hermiticity, since as mentioned above, all 2×2 unimodular Hermitian matrices represent pure boosts just as all 2×2 unimodular unitary matrices represent pure rotations.

As a last example to demonstrate the utility of the Pauli algebra, consider the inverse problem[16] of finding the boost parameter $\underline{\mathbf{w}}$, or equivalently, the corresponding four-velocity $\underline{u} = \exp(\underline{\mathbf{w}})$, in terms of a given initial four-vector, say $r = t\,\underline{1} + \underline{\mathbf{r}}$, and the given transformed result

$$r' = \underline{B}(\mathbf{w})\underline{r}\underline{B}^{+}(\mathbf{w}) = \underline{u}^{1/2}\underline{r}\underline{u}^{1/2} = \underline{\mathbf{r}}_{\perp} + \underline{u}\underline{r}_{\parallel}, \qquad (38)$$

where $\underline{r}_{\parallel} = t\underline{1} + \underline{\mathbf{r}}_{\parallel}$ and $\mathbf{r}_{\parallel} = \mathbf{r}\cdot\hat{u}\hat{u} = \mathbf{r} - \mathbf{r}_{\perp}$. Since by Eq. (38)

$$\underline{r}' - \underline{r} = (\underline{u} - \underline{1})\underline{r}_{\parallel}, \qquad (39)$$

The vector part of $r' - \underline{r}$ is parallel to $\mathbf{u}$ and consequently determines $\pm\,\hat{u}$, and hence also $\underline{r}'_{\parallel}$ and $\underline{r}_{\parallel}$. It is clear from Eq. (38) that $\underline{r}'_{\parallel} = \underline{u}\underline{r}_{\parallel}$ and therefore

$$\underline{u} = \underline{r}'_{\parallel}\underline{\bar{r}}_{\parallel}/\underline{r}_{\parallel}\underline{\bar{r}}_{\parallel}. \qquad (40)$$

We are able to divide by the matrix product $\underline{r}_{\parallel}\underline{\bar{r}}_{\parallel}$ because any 2×2 matrix times its spatial reverse is a scalar.

A similar solution exists when the given initial and boosted quantities are six-vectors[4] such as the electromagnetic field $\mathbf{F} = \mathbf{E} + i\mathbf{B}$, which transforms under boosts as follows [compare Eq. (17)]:

$$\underline{\mathbf{F}}' = \underline{B}(\mathbf{w})\underline{\mathbf{F}}\underline{B}^{-1}(\mathbf{w}) = \underline{u}^{1/2}\underline{\mathbf{F}}\underline{u}^{-1/2} = \mathbf{F}_{\parallel} + \underline{u}\underline{\mathbf{F}}_{\perp}. \qquad (41)$$

Since by Eq. (41)

$$\underline{\mathbf{F}}' - \underline{\mathbf{F}} = (\underline{u} - 1)\underline{\mathbf{F}}_{\perp}, \qquad (42)$$

$\pm\,\hat{u}$ is normal to the plane in which the real and imaginary parts of $\mathbf{F}' - \mathbf{F}$ lie. [Should $\mathbf{E}' - \mathbf{E}$ and $\mathbf{B}' - \mathbf{B}$ be parallel, one can show that $\pm\,\hat{u}$ is parallel to $(E' - E)(\mathbf{E}' + \mathbf{E}) + (B' - B)(\mathbf{B}' + \mathbf{B})$.] Then Eq. (41) gives $\underline{u}$ as

$$\underline{u} = \frac{(\underline{\mathbf{F}}' - \underline{\mathbf{F}}_{\parallel})\underline{\mathbf{F}}_{\perp}}{\mathbf{F}_{\perp}\cdot\mathbf{F}_{\perp}} = \frac{\underline{\mathbf{F}}'_{\perp}\underline{\mathbf{F}}_{\perp}}{\mathbf{F}_{\perp}\cdot\mathbf{F}_{\perp}}. \qquad (43)$$

V. CONCLUSIONS

We have shown that for most calculations in special relativity, the Pauli algebra $\mathscr{P}$ is fully as powerful as the Dirac algebra $\mathscr{D}$. The advantages of $\mathscr{P}$ over $\mathscr{D}$ are based on its relative simplicity: it has only half as many basis forms and all multiplication among elements can be expressed by familiar multiplication rules of the 2×2 Pauli spin matrices or, equivalently, can be expressed in terms of familiar dot and cross products of spatial vectors in three-dimensional Euclidean space. The fact that restricted Lorentz transformations in $\mathscr{P}$ take the simple $\mathrm{SL}(2,C)$ form is an added bonus.

Although the examples treated here have been restricted to relativistic kinematics, the power of $\mathscr{P}$ is even more evident in electrodynamics.[8] In $\mathscr{P}$, for example, Maxwell's four equations result trivially as the scalar, vector, pseudoscalar, and pseudovector parts of the field equation for the vector potential $\underline{A}$. Furthermore, the analytic solution describing the motion of a charge in arbitrary constant electric and magnetic fields is simpler to find in the Pauli algebra[19] than in the Dirac algebra, and the effect of the fields on the four-velocity is easily seen to induce a well-defined Lorentz transformation. General solutions for the motion of a charge in a circularly polarized plane wave are also easily derived in $\mathscr{P}$.[19]

Of course it may happen that extensions, perhaps to curved space-time geometries, may require a more complex covering algebra, perhaps $\mathscr{D}$ or even a higher-order Clifford algebra. Even then, the Pauli algebra should remain useful, not only for calculations in flat space-time, but also as an introduction to Clifford algebras, since it is the simplest nontrivial example of a Clifford algebra over the field of complex numbers.[7]

ACKNOWLEDGMENT

This work was supported by the Natural Sciences and Engineering Research Council of Canada.

[1] D. Hestenes, *Space-Time Algebra* (Gordon and Breach, New York, 1966).

[2] Y. Choquet-Bruhat, C. Dewitt-Morrette, and M. Dillard-Bleick, *Analysis, Manifolds, and Physics* (North-Holland, Amsterdam, 1977).

[3] N. Salingaros and M. Dresden, Adv. Appl. Math. **4**, 1, 31 (1983).

[4] N. Salingaros, J. Math. Phys. **25**, 706 (1984).

[5] N. Salingaros, J. Math. Phys. **27**, 157 (1986); **28**, 492 (E) (1987).

[6] N. Salingaros, Phys. Rev. D **31**, 3150 (1985).

[7] N. Salingaros, J. Math. Phys. **23**, 5 (1982).

[8] W. E. Baylis, Am. J. Phys. **48**, 918 (1980).

[9] C. Møller, *The Theory of Relativity* (Oxford U. P., London, 1972), 2nd ed.

[10] J. D. Jackson, *Classical Electrodynamics* (Wiley, New York, 1975), 2nd ed.

[11] The Dirac Hamiltonian combines vector (odd) and scalar (even) elements, but the Dirac equation can be easily written as two coupled two-component spinor equations involving only 2×2 matrices. See, for example, Ref. 8, Sec. V J.

[12] F. Klein and A. Sommerfeld, *Uber die Theorie des Kreisels* (Teubner, Leipzig, 1910), Vol. IV.

[13] For example, P. Weiss, Proc. R. Irish Acad. **46**, 129 (1941); C. W.

Misner, K. S. Thorne, and J. A. Wheeler, *Gravitation* (Freeman, San Francisco, 1973); and citations in Ref. 8.

[14] A. J. MacFarlane, J. Math. Phys. **3**, 1116 (1962).

[15] C. B. van Wyk, Am. J. Phys. **52**, 853 (1984).

[16] C. B. van Wyk, J. Math. Phys. **27**, 1311 (1986).

[17] N. Salingaros and Y. Ilamed, Found. Phys. **14**, 777 (1984).

[18] The spatial reversal of any element is formed by changing the sign of the spatial vector and pseudovector components. The notation for this is as in Ref. 4, but the expression "spatial inverse" used there would be better reserved for the Hermitian conjugate of the spatial reversal, in which the signs of vector and pseudoscalar parts are reversed while scalar and pseudovector components remain unaffected.

[19] W. E. Baylis and G. Jones, "Relativistic dynamics of charges in external fields: The Pauli algebra approach," to be published.

Solitary waves as fixed points of infinite-dimensional maps for an optical bistable ring cavity: Analysis

H. Adachihara, D. W. McLaughlin, J. V. Moloney, and A. C. Newell
Program in Applied Mathematics, University of Arizona, Tucson, Arizona 85721

(Received 5 February 1987; accepted for publication 15 July 1987)

The transverse behavior of a laser beam propagating through a bistable optical cavity is investigated analytically and numerically. Numerical experiments that study the (one-dimensional) transverse structure of the steady state profile are described. Mathematical descriptions of (i) an infinite-dimensional map that models the situation, (ii) the solitary waves that represent the transverse steady state structures, (iii) a projection formalism that reduces the infinite-dimensional map to a finite-dimensional one, and (iv) the theoretical analysis of this reduced map are presented in detail. The accuracy of this theoretical analysis is established by comparing its predictions to numerical observations.

I. INTRODUCTION

When one observes physical systems with a large number of degrees of freedom, one frequently notices robust configurations that remain spatially coherent even though the temporal evolution of the system is chaotic.[1] Vortices in a two-dimensional turbulent flow and cellular patterns in Benard convection are two examples of such coherent phenomena. For large-dimensional systems, which are usually described by nonlinear partial differential equations, it is extremely difficult to describe the spatially coherent, yet temporally chaotic, solutions, theoretically and virtually impossible at present to capture solutions that are both temporally and spatially chaotic. In this series of papers we will study an important physical problem—that of a laser beam propagating through a bistable optical cavity—which admits a rather complete theoretical description of its chaotic states. This problem is intriguing in that it is both technologically important in laser optics (for the development of an all optical means to process signals such as would be needed in an all optical computer[2]) and theoretically important in turbulence[1] (as a tractable example of a spatially coherent, temporally chaotic, nonlinear field).

Our goals in this first paper of the series are as follows.

(1) *To describe in detail the results of our numerical experiments* in which we observe the (one-dimensional) transverse structure of steady state fixed points of the laser beam profile.

(2) *To present self-contained mathematical descriptions of* (i) an *infinite-dimensional map* that models the situation, (ii) the *solitary waves* that represent the transverse structures, (iii) a *projection formalism* that reduces the infinite-dimensional map to a finite-dimensional one, and (iv) the *theoretical analysis of this reduced map.*

(3) *To establish the accuracy of this theoretical analysis* by comparing its predictions to the numerical observations. In the conclusion of the paper we will indicate typical routes the system takes to chaos as its fixed points become unstable. However, detailed descriptions of these routes to chaos will be deferred to a later paper in the series.

The paper is organized as follows: In Sec. II, we describe the infinite-dimensional map. This map is derived, from the Maxwell–Bloch equations, in Appendix A. In Sec. III, we summarize properties of the map in the plane wave case. The simple material in this section orients our entire study. In Sec. IV, we describe in detail our numerical experiments in which we investigated the (one-dimensional) transverse structure of the field as it emerges from the optical ring cavity. In Sec. V, we (i) define solitary waves, (ii) establish numerically that the fixed points profile contains solitary waves, (iii) develop a solitary wave projection formalism, and (iv) use this projection formalism to reduce the infinite-dimensional map to a finite-dimensional one. Details of the projection formalism are presented in Appendix B. In Sec. VI, we analyze the reduced map on solitary wave parameters, both analytically and numerically. In Sec. VII, we present an analytical formula with sufficient generality to fit the entire single fixed point, both its solitary wave central peak and its shape in the outer regions (henceforth referred to as wings). Finally, in the conclusion, Sec. VIII, we briefly describe routes of the system to chaos as the system is further stressed and the fixed points have become unstable. Some of the results discussed within were reported earlier,[1,3–5] together with our related works.[6–8]

II. DEFINITION OF THE MAP

The mathematical problem we study in this paper can be stated concisely. We study an infinite-dimensional map,

$$G_{n-1}(\cdot,l) \rightarrow G_n(\cdot,l), \tag{2.1}$$

for a function $G_n(x,z=l)$ defined by a sequence of initial value problems:

$$2i\frac{\partial}{\partial z}G_n + \frac{1}{f}\frac{\partial^2}{\partial x^2}G_n + N(G_nG_n^*)G_n = 0, \tag{2.2a}$$

$$G_n(x,z=0) = a(x) + Re^{i\phi}G_{n-1}(x,z=l). \tag{2.2b}$$

Here $n \geqslant 1$, $G_0 = 0$, (f,R,ϕ,l) are given real parameters, and $a(x)$ is a given function shaped like a Gaussian with maximum $A = a(0)$. Our goal is to find the behavior of the function $G_n(x,l)$ as $n \rightarrow \infty$.

The physical origin of this mathematical problem is as

0022-2488/88/010063-23$02.50

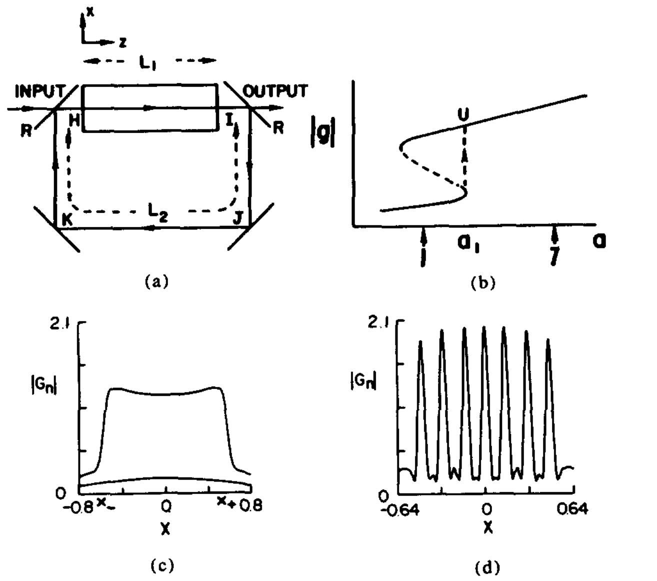

FIG. 1. Schematic of an optical ring resonator. A laser beam (Gaussian in x) enters through a partially transmitting mirror H, propagates a distance L_1 through a nonlinear dielectric medium, is partially transmitted for detection through the output mirror I while the remainder is fed back around the circuit (via mirrors I, J, K, and H) and added to the input beam. Mirrors J and K are 100% reflecting and serve only to direct the beam around the closed loop. For the purposes of mathematical convenience in the present paper, the return circuit is approximated by a simple linear phase shift.

follows: Consider a ring cavity as depicted in Fig. 1. A laser beam enters this cavity, which is filled with a nonlinear dielectric, emerges from it, and is brought back to the entry point by a rectangular array of four mirrors. There it reinforces the pump field and together they reenter the nonlinear medium. The goal is to predict the output field after many passes through the cavity.

A sketch of the derivation of the mathematical model (2.2) is given in Ref. 3; a more detailed derivation is given in Appendix A together with references to the physical literature. In the mathematical model, (2.2a) describes the propagation of the laser field down the nonlinear medium, while (2.2b) describes the return of the field to the reentry point. Here G_n denotes the envelope of the electromagnetic field on the nth pass through the cavity, $a(x)$ is the envelope of the input laser field, and x and z are coordinates in directions transverse and parallel to the direction of propagation through the medium. The parameter f is related to the Fresnel number which measures the amount of transverse dispersion or diffraction in propagation through the medium. The dynamics of the return path is accounted for by the factor $Re^{i\phi}$, which involves mirror losses ($R<1$) and a phase shift ϕ.

In this paper we study two nonlinear media, one with "saturable nonlinearity"

$$N(GG^*) = -1/(1+2GG^*), \tag{2.3a}$$

and the other with "Kerr nonlinearity"

$$N(GG^*) = -1+2GG^*. \tag{2.3b}$$

The latter is the first two terms in the Taylor series expansion of the former for small values of GG^*.

The infinite-dimensional map (2.1) is the composition of two maps—one, (2.2b), is a dissipative discrete map that acts pointwise in x on the output $G_{n-1}(x,l)$ to produce the input $G_n(x,0)$, while the second, (2.2a), is a conservative nonlinear wave equation that transports the field down the nonlinear medium. In our analysis we will utilize separately properties of each of these two components of map (2.1). For example, the nonlinear wave equation component will filter the field into coherent spatial structures, while the discrete dissipative map will select specific values for the amplitudes of these spatial structures.

III. PLANE WAVE CASE

In this section we begin our analytical study of map (2.2) by assuming the input field $a(x)$, and therefore $G_n(x,z)$, is independent of x. In this case the infinite-dimensional map reduces to a two-dimensional one which is quite tractable. We will use this reduced, two-dimensional map to orient our study. (This x-independent situation is called the "plane wave case" because the wave fronts are planar, with no transverse curvature; it has been studied by many authors, those in Ref. 9.)

When G_n is independent of x, the partial differential equation (2.2a) reduces to an ordinary differential equation that can be integrated explicitly. We find

$$G_n(z) = \exp[(i/2)N(G_nG_n^*)z]G_n(0). \tag{3.1}$$

Notice that $G_nG_n^*$ is constant, independent of z. Writing $g_n = G_n(0)$ and inserting (3.1) into (2.2b) yields the map

$$g_{n+1} = T(g_n,g_n^*) = a + R\exp[i(\phi + N(g_ng_n^*)l/2)]g_n. \tag{3.2}$$

The plane wave map (3.2) is a one-dimensional complex, or a two-dimensional real, invertible map. Its inverse is given explicitly by

$$\begin{aligned} g_n &= T^{-1}(g_{n+1},g_{n+1}^*) \\ &= \exp\left\{-i\left[\phi + N\left(\frac{(g_{n+1}-a)(g_{n+1}^*-a)}{R^2}\right)\frac{l}{2}\right]\right\} \\ &\quad\times(g_{n+1}-a)/R. \end{aligned} \tag{3.3}$$

Since the map T depends upon $g_ng_n^*$, it is not analytic in g_n. The map T works as follows (Fig. 2): First, the point $g=g_n$ is rotated through an angle $\theta = (\phi + N(gg^*)l/2)$ that depends nonlinearly upon g; then the new point is moved along a ray toward the origin by a contraction factor $R<1$; finally, the point is shifted toward the right by a positive real number a. This geometrical description of the action of the map T, as depicted in Fig. 2, makes the existence of fixed points plausible.

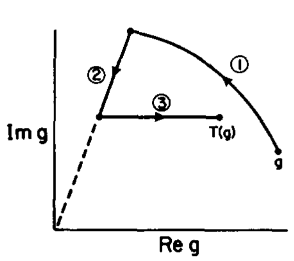

FIG. 2. Complex g-plane sketch showing that the action of the plane wave map is a composition of a ① rotation, ② contraction, and ③ translation [see Eq. (3.2)].

The following physical argument also makes the existence of fixed points plausible. Dissipation is present in the form of losses at the mirrors [which, by the way, is the only loss mechanism modeled in map (2.2)]. Thus we expect fixed points to result from a balance between energy lost at the mirrors and energy injected through the constant pump field "a."

Mathematically, the presence of losses is seen by a calculation of the Jacobian of map T:

$$|D_g T| = \det \begin{bmatrix} \dfrac{\partial T_R}{\partial g_R} & \dfrac{\partial T_I}{\partial g_R} \\ \dfrac{\partial T_R}{\partial g_I} & \dfrac{\partial T_I}{\partial g_I} \end{bmatrix} = \frac{1}{2i|g|} \det \begin{bmatrix} T_\theta & T_\theta^* \\ T_{|g|} & T_{|g|}^* \end{bmatrix}$$

$$= R^2 < 1, \tag{3.4}$$

where $T = T_R + iT_I, g = g_R + ig_I = |g|\exp(i\theta)$. Note that $|D_g T| = R^2 < 1$ is constant in g and dissipative.

Fixed points of T do exist. A fixed point g satisfies

$$g = T(g,g^*) = a + R\exp[i(\phi + N(gg^*)l/2)]g. \tag{3.5}$$

This equation can be analyzed as follows:

$$\Gamma = \Gamma(gg^*) \equiv \phi + N(gg^*)l/2, \tag{3.6a}$$

$$1 - \frac{a}{g} = Re^{i\Gamma} \Rightarrow \begin{cases} (1 - a/g)(1 - a/g^*) = R^2, \\ 2 - a(g + g^*)/gg^* = 2R\cos\Gamma. \end{cases}$$

These two equations quickly yield a single equation for gg^*,

$$\cos\Gamma(gg^*) = \tfrac{1}{2}(1/R + R - a^2/Rgg^*). \tag{3.6b}$$

Using a root gg^* of (3.6b) yields the fixed point

$$g = a/(1 - Re^{i\Gamma}). \tag{3.6c}$$

Thus the fixed points are determined by (3.6b). These arise intuitively as follows: In the absence of nonlinearity and mirror losses, the cavity is described by the map

$$g_{n+1} = a + \exp[i(\phi - l/2)]g_n.$$

The resonant response of this empty cavity is given by the condition $\exp[i(\phi - l/2)] = 1$; Eqs. (3.6a) and (3.6b) constitute the generalization of this phase match condition after mirror loss and nonlinearity have been added to the system. Qualitative insight into (3.6b) may be obtained graphically. It is easiest to begin with the Kerr nonlinearity:

$$I = gg^*, \quad N(I) = -1 + 2I, \quad \Gamma(I) = \phi - l/2 + lI,$$
$$\cos\Gamma = \tfrac{1}{2}(1/R + R - a^2/RI).$$

The situation is sketched in Fig. 3. Figure 3 shows several points.

(i) The parameter l, a dimensionless measure of the length of the nonlinear medium, primarily determines the number of fixed points which increases as l increases.

(ii) In the situation depicted, there are three fixed points. For smaller input intensity a^2, the curve $(R^{-1} + R - a^2/RI)$ rises faster and only one fixed point exists. For larger values of a^2, the same curve falls and only one fixed point remains. Thus the fixed points as function of a^2 take on the familiar "hysteresis" shape of Fig. 4.

(iii) The maximum possible switch-up intensity is bounded by

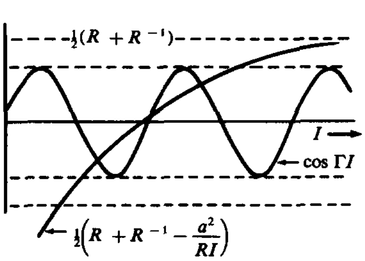

FIG. 3. Graphical construction to determine fixed points of the plane wave Kerr map. For the case shown there are three fixed points located at the intersection points of the two curves.

$$I_{\text{Max}} - I_{\text{Min}} = \frac{a^2}{(1-R)^2} - \frac{a^2}{(1+R)^2}$$

$$= \frac{4R}{(1-R)^2(1+R)^2}a^2. \tag{3.7}$$

While this switch-up is linear in a^2, there is the possibility of substantial gain for small mirror losses.

For other than cubic nonlinearities, very similar properties hold. For example, consider the saturable nonlinearity $N(I) = -(1 + 2I)^{-1}$. A good way to picture the situation is as follows: First, sketch the curve $(R + R^{-1} - a^2/RI)/2$ [see Fig. 5(a)]. Then on a $\cos\Gamma$ curve find ϕ and $\phi - l$, $(\phi > l)$ [see Fig. 5(b)]. Finally, use the monotonic transformation (3.6a) $\Gamma(I) = \phi - [1/(1 + 2I)]l$ to stretch the $\cos(\cdot)$ curve over the entire I axis $(\phi - l \simeq I = 0, \phi \simeq I = +\infty)$ [see Fig. 5(c)]. From this construction one sees that the qualitative features describe for the Kerr fixed points persist for the saturable nonlinearity. In particular, one again obtains a hysteresis curve like Fig. 4. To obtain more quantitative information about the fixed points, the transcendental equation (3.6b) must be solved numerically. Next, we turn to a discussion of the stability of these fixed points.

Consider the map $T(\cdot)$,

$$h = T(f,f^*) = a + R\exp[i(\phi + N(ff^*)l/2)]f, \tag{3.8}$$

and linearize it about a fixed point g. That is, write $f = g + \tilde{f}$, $h = g + \tilde{h}$, and retain only linear terms in $\tilde{f},\tilde{h}$:

$$\begin{pmatrix} \tilde{h} \\ \tilde{h}^* \end{pmatrix} = D_g T \begin{pmatrix} \tilde{f} \\ \tilde{f}^* \end{pmatrix}, \tag{3.9a}$$

where $D_g T$ denotes the derivative of T at (g,g^*),

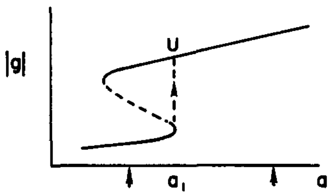

FIG. 4. Hysteresis (bistable) response curve for the case represented by three fixed points (see Fig. 3).

$$D_g T = \begin{bmatrix} \partial_f T & \partial_{f^*} T \\ \partial_f T^* & \partial_{f^*} T^* \end{bmatrix} \Bigg|_{\substack{f=g,\\ f^*=g^*}} = \begin{bmatrix} (1 + i(l/2)N'(gg^*)gg^*)Re^{i\Gamma} & i(l/2)N'(gg^*)g^2 e^{i\Gamma}R \\ -i(l/2)N'(gg^*)g^{*2}e^{-i\Gamma}R & (1 - i(l/2)N'(gg^*)gg^*)Re^{-i\Gamma} \end{bmatrix}. \tag{3.9b}$$

We recall that the Jacobian of $T(g,g^*) = \det D_g T = R^2$; also, note that if $(\mu, \mathbf{v})$ denotes an eigenvalue, eigenvector pair of $D_g T$, then so does $(\mu^*, \mathbf{w} = (v_2^*, v_1^*))$. The eigenvalues (μ_1,μ_2) of $D_g T$ satisfy

$$\mu_1\mu_2 = R^2 < 1, \tag{3.10a}$$

$$\mu_1 + \mu_2 = [1 + i(l/2)N'(gg^*)gg^*]Re^{i\Gamma} + [1 - i(l/2)N'(gg^*)gg^*]Re^{-i\Gamma}. \tag{3.10b}$$

The fixed point g is linearly stable to forward iterations if and only if $|\mu_j| \leqslant 1$, for $j = 1,2$. Here two cases arise: (i) μ_1 and μ_2 both real, or (ii) $\mu_2 = \mu_1^*$. The second case of conjugate pairs is always stable, since $\mu_1\mu_2 = \mu_1\mu_1^* = R^2 < 1$. Thus no "Hopf bifurcations" are allowed. In the first case of real eigenvalues, either both satisfy $|\mu_j| < 1$, or one satisfies it and the other does not. As an eigenvalue μ crosses $+1$, the instability which occurs is a "saddle-node" bifurcation. As an eigenvalue μ crosses -1, a period-2 bifurcation occurs. Typical trajectories of the μ eigenvalues as the parameters change are depicted in Fig. 6.

This linearized stability analysis can be applied to hysteresis loops such as Fig. 4. In many instances, for lower values of the stress parameters a and l, one finds that the upper and lower branches are stable, while the intermediate branch is unstable. Thus we have the necessary ingredients for an on–off optical switch.[2] However, as the strengths of the parameters are increased, branches can destabilize to period doubling bifurcations although, in the case of the saturable nonlinearity, the upper most branch appears to be always stable.

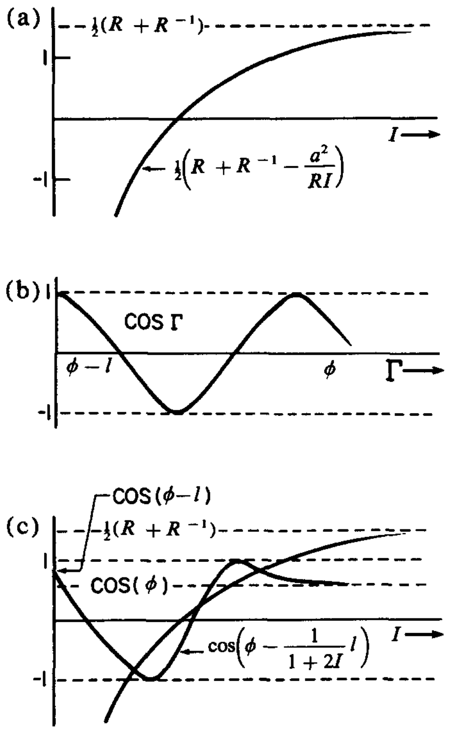

FIG. 5. Similar graphical fixed point determination to that shown in Fig. 3. Here, however, the nonlinearity is saturable. (a) and (b) The right and left sides of the equality (3.6b); (c) the superposition of both.

Rather than continue the detailed algebraic analysis of (3.10) in order to study the linearized stability, we iterate T numerically. These studies are summarized in Fig. 7, which shows a bifurcation diagram as a function of the parameters l and a^2. This diagram shows stable fixed points going unstable to period-2 states, which in turn go unstable through a period doubling route to chaos. It also shows that different chaotic states (with different basins of attraction) can coexist at the same parameter values.

Numerical experiments can be used to study these chaotic attractors. For example, a sequence of iterates can be plotted on the complex phase plane, Fig. 8. This figure shows the leafy-Cantor-like structure which is now familiar in "strange attractors."

Using global phase space methods,[8] considerably more information can be obtained about the map T. In particular, it is important to understand the universal, self-similar properties of this two-dimensional, invertible map in a class of maps which are the composites of a nonlinear rotation, a contraction, and a shift. However, for our purposes the more elementary analysis summarized here is sufficient.

We close this section on the plane wave map by returning the reader's attention to its hysteresis diagram, Fig. 4, which will play an important role in the next section.

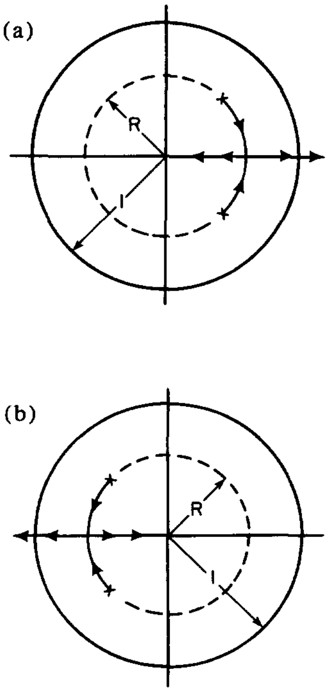

FIG. 6. A sketch of the behavior of the plane wave map eigenvalue pair (μ_1,μ_2) near a (a) saddle node and (b) period doubling bifurcation.

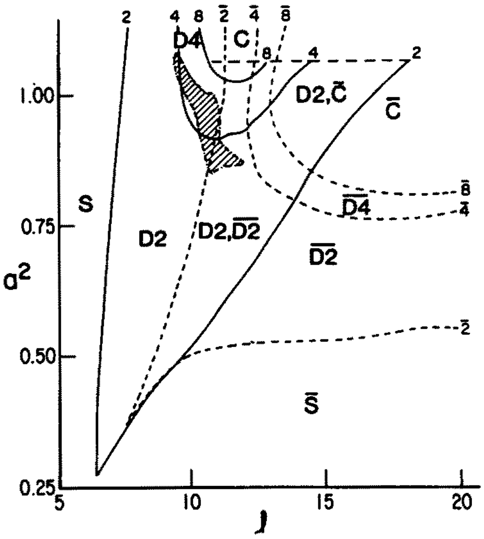

FIG. 7. Numerically generated bifurcation diagram in coordinates (a^2,l) for the plane wave map [Eq. (3.2)] (Ref. 8a). This diagram shows only a part of the full picture discussed in Ref. 8b. Two distinct period doubling routes to chaotic attractors are indicated by the unbarred and barred symbols respectively. [S($\bar{S}$)—stable fixed point, D2($\overline{D2}$)—period 2, D4($\overline{D4}$)—period 4, C($\bar{C}$)—chaotic attractors]. These attractors coexist in wide regions of parameter space. The cross-hatched region contains a period-6 attractor that can coexist with attractors from either route (Ref. 8a).

IV. NUMERICAL DESCRIPTION OF FIXED POINTS WITH TRANSVERSE STRUCTURES

We turn to the original problem (2.2), with one-dimensional transverse effects included by considering an input field $a(x)$ which has a Gaussian-like transverse profile,

$$2i\frac{\partial}{\partial z}G_n+\frac{1}{f}\frac{\partial^2}{\partial x^2}G_n+N(G_nG_n^*)G_n=0, \qquad (4.1a)$$

$$G_n(x,0)=a(x)+Re^{i\phi}G_{n-1}(x,l). \qquad (4.1b)$$

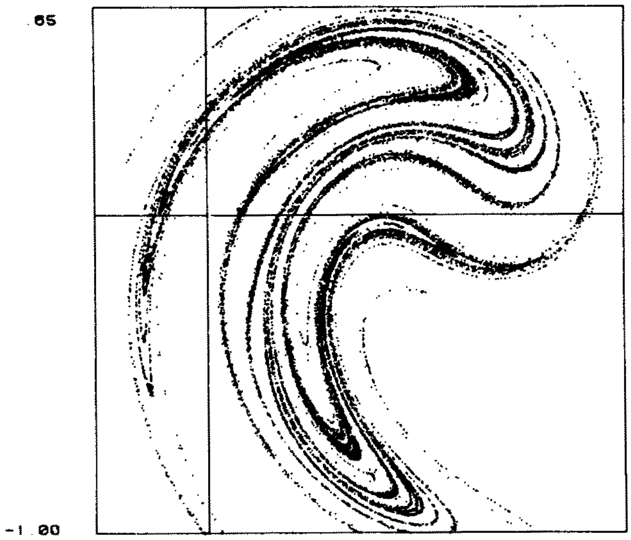

FIG. 8. Chaotic attractor numerically generated by iterating the plane wave map many thousands of times for a fixed set of parameter values in region $\bar{C}$ of Fig. 7.

Clearly this infinite-dimensional map possesses a wide variety of potential responses depending upon parameter values. We restrict these by focusing our attention (for the most part) on large Fresnel numbers $[F=(\ln 2/4\pi)f=100\gg 1]$ and selecting the parameters ϕ, l, and R in regions where the plane wave map has a hysteresis diagram such as Fig. 4.

In Eq. (4.1a) the only coupling of a transverse segment of the beam profile to its neighbors occurs through the Laplacian $f^{-1}\partial_{xx}$. For large Fresnel numbers and moderate transverse gradients in the input pulse, $f^{-1}\partial_{xx}a(x)\ll a(x)$ and this coupling can be initially neglected. Initially, then, each transverse segment of the profile acts independently from its neighbors and its dynamic are determined by a local plane wave theory. Thus those points on the Gaussian profile for which $a(x)>a(x_\pm)=\sqrt{I_1}$ (see Fig. 4) will switch up to the upper branch while those parts for which $a(x)<a(x_\pm)=\sqrt{I_1}$ will go to the lower branch. The center of the beam profile will switch up, while its wings will not. For the saturable nonlinearity, with parameter values set at $[F=100, l=2, \phi=0.4, |a(0)|^2=0.0375]$, this situation is shown in Fig. 9. The Fresnel number F is related to f by $F=[(\ln 2)/4\pi]f$. In Fig. 9 we show that initial Gaussian profile, and the output profile after 23 passes through the nonlinear medium. Notice that the center of the profile has switched to the upper branch while the wings have switched to the lower branch according to the plane wave theory.

However, now the two outer edges of the profile possess a steep gradient and, near $x\simeq x_\pm$, the term $f^{-1}\partial_{xx}G_n$ is no longer negligible and the plane wave approximation is no longer valid. Numerical experiments (which solve the full partial differential equation) show what happen during this stage of the evolution. At the edges $x_\pm$ narrow spatial rings of width $\Delta x=O(1/\sqrt{f})$ are generated (Fig. 10). These narrow rings eventually fill out the region between x_- and x_+. Once these spatial rings form and fill up the transverse profile, they persist and describe the large n asymptotic response of the infinite-dimensional map. These rings can become the steady state response of the system; that is, they can be stable fixed points of the infinite-dimensional map.

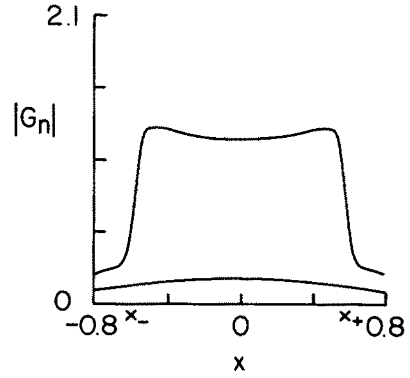

FIG. 9. Switched-on transverse beam profile $G_n(x,l)$ at iterate $n=20$ of the infinite-dimensional map [Eqs. (4.1a) and (4.1b)]. The sharp gradients at $x_\pm$ induce strong local transverse coupling and should be contrasted with the smooth initial pump profile $a(x)$ also shown in this figure. Parameters used to generate this and Fig. 10: $F=100$, $l=2$, $\phi=0.4$, $|a(0)|^2=0.0375$.

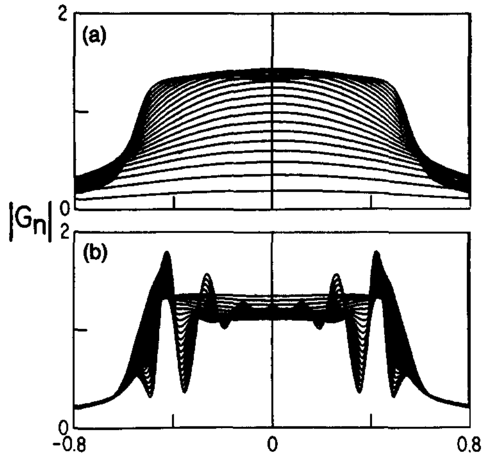

FIG. 10. Transient switching of the transverse profile $G_n(x,l)$ from the initially broad and smooth pump profile $a(x)$. (a) The development of the sharp gradient at the edges ($x_\pm$) and saturation at beam center over the first 20 iterates of Eqs. (4.1a) and (4.1b). The initiation of the transverse spatial rings occurs at these sharp edges. These rings grow inwards towards beam center. (b) Iterates $n = 20$–40.

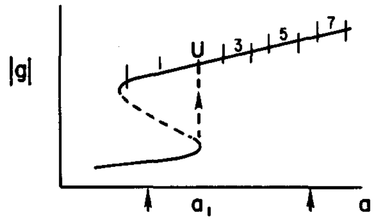

FIG. 12. Hysteresis loop showing the relative disposition of stationary n-ring ($n = 1,3,5,7$) fixed points on the upper branch of the loop. In the gaps between the marked-off regions, the transverse profile never settles down to a stationary ring pattern but instead undergoes a complicated oscillatory motion as discussed in the text.

For example, a fixed point does emerge from the transient pictured in Fig. 9. This fixed point, as shown in Fig. 11 at the 200th resonator pass, is a seven stationary ring pattern whose rings sit on a broad background or shelf which can be identified with the lower branch fixed point of the plane wave case.

The number of rings can be controlled. For the saturable nonlinearity, we have observed 1, 3, 5, 7 stationary rings. The actual number of rings seems to be a function of the transient shape realized after ~20 passes; it is determined approximately by the number of rings of width $1/\sqrt{f}$ which can fit in that portion of the beam which switches to the upper branch, i.e., number of rings $\alpha\sqrt{f}(x_+ - x_-)$. Our numerical calculations show that the upper branch of the hysteresis curve can be segmented into small bands in $|a(0)|^2$, where n spatial rings, with n an odd interger, are the stable steady states of the system. For example, the experiment with parameter values set at ($F = 100$, $l = 2$, $\phi = 0.4$), the segment $0.008 < |a(0)|^2 < 0.018$ has a fixed point which consists of a single "ring," whose spatial profile has an amplitude which increases and a width ($\sim 1/\sqrt{f}$) which decreases as $|a(0)|^2$ increases from 0.008 to 0.018. Increasing $|a(0)|^2$ further, for example, to $|a(0)|^2 = 0.025$, so that it lies well beyond the switch-up point (a in Fig. 4) produces a stationary three-ring structure. Between the $n = 1$ and $n = 3$ stationary ring regions there exists a finite range of $|a(0)|^2$ where one observes a slow recurrent oscillation between one and two or two and three ring patterns. Even numbers of spatial rings can appear initially, but they continue to oscillate in a complicated manner on further circuits of the resonator. Whether the rings become stationary or not seems to depend on whether an odd integral number rings of width $\sim 1/\sqrt{f}$ can fit into the total area of the switched-on portion of the beam (see Fig. 9). The relative disposition of the n-ring stationary transverse spatial structures on the upper branch of the hysteresis is summarized in Fig. 12.

We have carried out the following numerical experiment to establish the role of the external pump and dissipation [Eq. (2.2b)] in stabilizing these transverse ring structures. The stationary rings that developed after 200 resonator passes were taken as initial values to the nonlinear evolution equation (2.2a) and propagated down a long tube. After a short distance, of the order of a few medium lengths in the resonator, the rings were observed to oscillate up and down in amplitude about their original stationary values. This demonstrates immediately that the map (2.2a) and (2.2b) acts to freeze out these rings structures. For three rings at $|a(0)|^2 = 0.025$ ($F = 100$, $l = 2$, $\phi = 0.4$) we observed the following behavior. Initially the rings appear to oscillate up and down but do not attain any noticeable velocities. They appear to be trapped by the broad shelf which represents the plane wave lower branch fixed point. On further propagation the shelf becomes modulated and low amplitude rings develop. Once these are well formed the two large amplitude outer rings in the triplet begin to slowly propagate outwards, interacting nonlinearly with and passing through their low amplitude neighbors. This behavior is reminiscent of soliton propagation. We tracked the evolution until both outer rings reached the "$1/e$" value

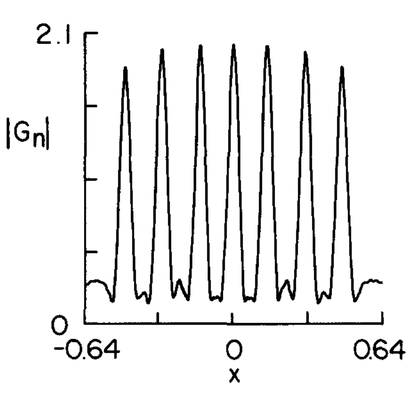

FIG. 11. This figure shows the large n ($n = 200$) asymptotic state of the map which follows the dynamic evolution of the preceding figure. This seven-ring stationary pattern represents a fixed point of the infinite-dimensional map [Eqs. (4.1a) and (4.1b)].

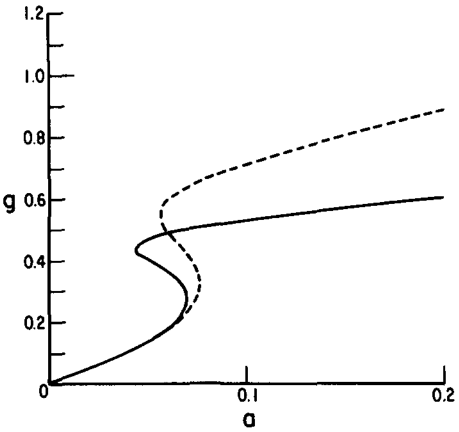

FIG. 13. Plane wave hysteresis loops for a Kerr (dashed) and saturable (solid line) nonlinearity. The parameters used to compute these curves are $l = 2, \phi = 0.6$.

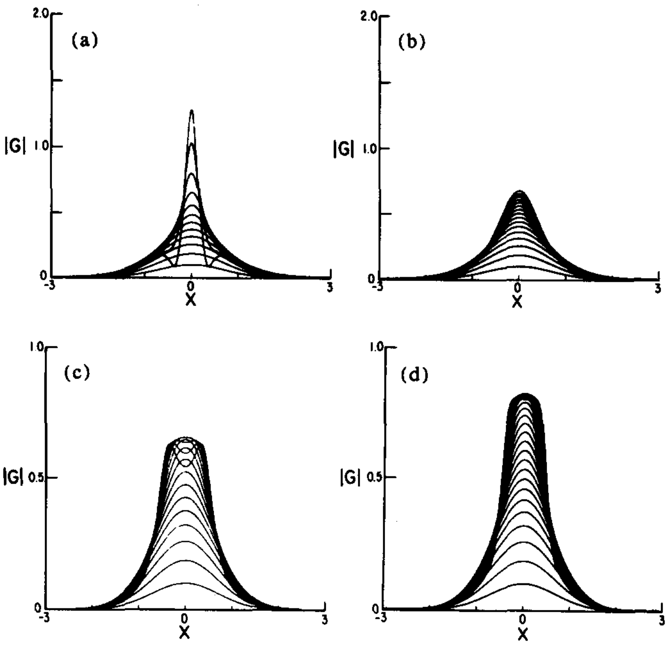

FIG. 15. Four cases showing detailed switch-on behavior of the beam $G_n(x,l)$. Cases (a) and (b) correspond to Kerr and saturable nonlinearities with $F = 1$, while (c) and (d) represent the same nonlinearities with $F = 100$. Parameters used are given in the text.

$(x = \pm 1)$ of the original input Gaussian beam. The center ring just oscillates up and down. These experiments indicate that the rings are independent structures entrained as a consequence of the pump and dissipation terms in the map.

As mentioned earlier, the rings sit on a broad background which, since f is large, should have the height of the plane wave lower branch fixed point. To verify this, we consider the experiment with parameter values $[F = 100, |a(0)|^2 = 0.008, l = 2, \phi = 0.6]$. For these values of l and ϕ, the plane wave hysteresis curves for both the saturable and Kerr nonlinearities are shown in Fig. 13. Using these hysteresis curves, we compute how the wings of the transverse profile should behave. The prediction is compared with the actual transverse experiments for the saturable case in Fig. 14 and shows perfect agreement.

The switch-up mechanism is primarily controlled by the sizes of the Fresnel number F. This is illustrated in Fig. 15 for both the saturable and Kerr nonlinearities. For this figure, the parameters $[l = 2, \phi = 0.6$, and $|a(0)|^2 = 0.01]$ were chosen, and for each nonlinearity, two experiments—one at $F = 1$ and a second at $F = 100$—were conducted. Notice the absence of a flattop in the beam profiles for $F = 1$. Although the switch-up procedure is predominantly determined by the size of F and is very similar for both nonlinearities, Fig. 15 does indicate that the beam grows in a somewhat more controlled manner in the saturable case.

The Fresnel number F also plays a major role in determining the width of the upper branch rings. When F is large, these rings are skinny and tall. For example, with $F = 100$, $|a(0)|^2 = 0.008, l = 2, \phi = 0.4$, the situation is as shown in Fig. 16.

We close this section by describing some of the differ-

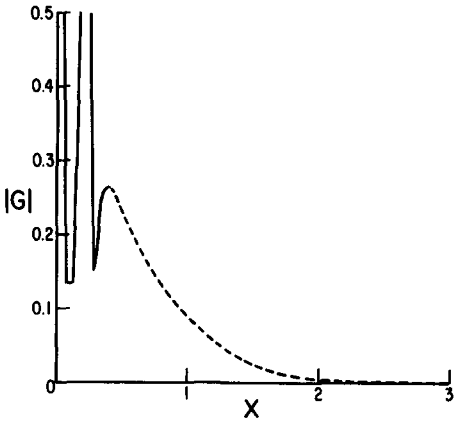

FIG. 14. A plot of one-half of the spatial three-ring fixed point with the predicted plane wave fixed point curve (dashed) superimposed on the wing. The transverse wing and plane wave fixed point curve are indistinguishable.

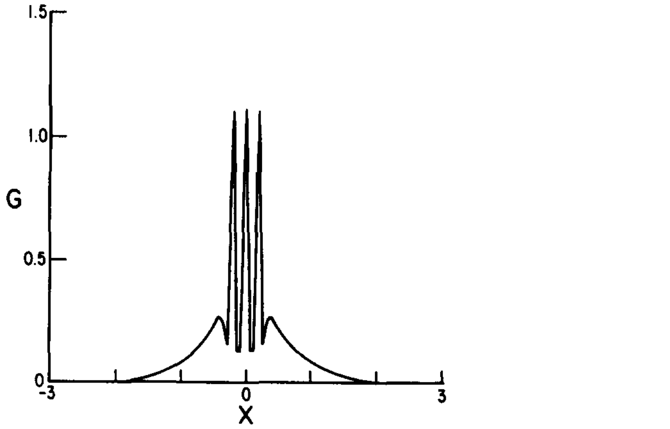

FIG. 16. An example of a three-ring fixed point showing the narrow rings sitting on a broad shelf corresponding to the lower branch plane wave fixed points (see Fig. 14).

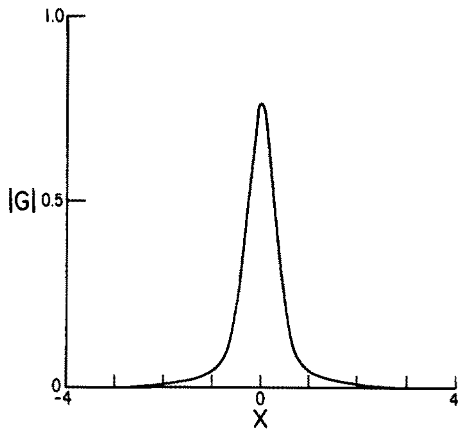

FIG. 17. An example of a single-ring Kerr fixed point at parameter values $F=0.8$, $l=1.5$, $\phi=0.4$, and $a(0)^2=0.003$.

ences that we have observed when the saturable nonlinearity is replaced by the Kerr nonlinearity. Generally, both the evolution of the beam and the final asymptotic states are much more sensitive to the pumping amplitude in the Kerr case. First, it is difficult to achieve a single-ring asymptotic state in the Kerr case. The only region on the upper branch where a single-ring shape appears to arise is at the extreme left near the switch down point. In Fig. 17 we show one such single-ring profile for $[F=0.8,\ l=1.5,\ \phi=0.4,\ |a(0)|^2=0.003]$. Second, fixed points are more difficult to achieve. As $|a(0)|^2$ is increased, extra rings do appear, but the beam approaches an oscillatory state rather than a fixed point. This state resembles an exact multisoliton solution of the integrable (Kerr) nonlinear Schrödinger equation. It should be contrasted to the multiring fixed points which arise in the saturable case and which appear to be phase-locked individual entities. To emphasize the distinctions we propagated both asymptotic states down an extended tube. In the Kerr case, the asymptotic state of the resonator persisted as a coherent transverse structure which did resemble (in one case at least) an analytically generated two-soliton waveform. On the other hand, in the saturable case, the individual rings did not remain locked together as they propagated down the long tube; instead, they drifted apart at different velocities.

We can conclude from these numerical studies that the asymptotic dynamical states for the field on the upper hysteresis branch differ significantly for saturable and Kerr nonlinearities. In the Kerr case, fixed points are rare; the asymptotic states then to be oscillatory. They also depend rather sensitively on the pumping amplitude. The saturable case has more controlled, more stable asymptotic responses.

V. SOLITARY WAVE REDUCTION OF THE MAP

A. Solitary waves

Known theory for the nonlinear wave equation (4.1a) indicates that the transverse rings should be solitary waves provided the nonlinear medium is sufficiently long. (Solitary waves are the asymptotic states of the propagation equation.[10]) A particular solitary wave is a solution of (4.1a) in the form $(y=\sqrt{f}x)$

$$G_s(y,z;\lambda)=S(\lambda y;\lambda)e^{i[(\lambda^2-1)z/2]}, \tag{5.1}$$

where $S(\theta;\lambda)$ is a real, even solution of

$$S_{\theta\theta}-S+(1/\lambda^2)[1+N(S^2)]S=0, \tag{5.2}$$

which vanishes as $\theta\rightarrow\infty$. The general solitary wave is a four-parameter family of solutions of (4.1a),

$$G_s(y,z;\lambda,\gamma,a,v)$$
$$=S[\lambda(y-a-vz);\lambda]e^{i[vy+(\lambda^2-1-v^2)z/2+\gamma]}, \tag{5.3}$$

which can be obtained form the particular solution (5.1) by using the symmetries of phase, translation, and Galilean invariance. In this work it will be sufficient to consider the two-parameter family

$$G_s(y,z;\lambda,\gamma)=S(\lambda y;\lambda)e^{i[(\lambda^2-1)(z/2)+\gamma]}, \tag{5.4}$$

because the transverse profiles are symmetric about the beam axis. (We prove *a posteriori* that this symmetry remains unbroken when the stresses are applied to this cavity.) The parameter λ determines the amplitude and width of the solitary wave, while γ determines its phase.

In the case of the Kerr nonlinearity $[1+N(S^2)=2S^2]$, the solitary wave takes the explicit form

$$S(\theta;\lambda)=\lambda\ \text{sech}\ \theta,\quad \theta=\lambda y. \tag{5.5}$$

Here λ certainly determines the amplitude and width of the solitary wave. For more general nonlinearities, one must study the differential equation (5.2). We illustrate for the saturable case where

$$1+N(S^2)=1-1/(1+2S^2)=2S^2/(1+2S^2). \tag{5.6a}$$

Introducing a potential $V(S)$, which for this saturable case takes the form

$$V(S)=\left(\frac{1}{\lambda^2}-1\right)\frac{S^2}{2}-\frac{1}{4\lambda^2}\ln(1+2S^2), \tag{5.6b}$$

the differential equation (5.2) may be rewritten as

$$S''=-\frac{\partial}{\partial S}V(S). \tag{5.6c}$$

This equation has the immediate "energy integral"

$$\tfrac{1}{2}S'^2=E-V(S). \tag{5.6d}$$

The potential $V(S)$ is sketched in Fig. 18. From this sketch of V, we see that E must be chosen as zero if we are to satisfy

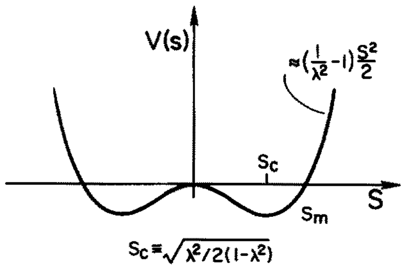
V(s)
$\approx(\frac{1}{\lambda^2}-1)\frac{S^2}{2}$
Sc
S
Sm
$S_c\equiv\sqrt{\lambda^2/2(1-\lambda^2)}$

FIG. 18. Sketch of the potential $V(s)$ for the saturable nonlinearity.

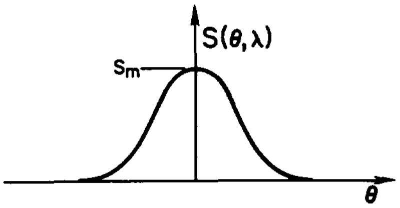

FIG. 19. Sketch of the solitary wave shape derived from the potential in Fig. 18.

the boundary condition at $\theta = \infty$, and that the parameter λ may take any value in the range $0 < \lambda < 1$. (No solitary wave exists for $\lambda > 1$.) With these considerations the qualitative shape of the solitary wave $S(\theta;\lambda)$ is sketched in Fig. 19. The amplitude $S_m = S_m(\lambda)$ is determined from

$$(1-\lambda^2)S_m^2 = \tfrac{1}{2}\ln(1+2S_m), \tag{5.6e}$$

and is a monotone increasing function of λ as λ runs from 0 to 1.

B. Numerical comparison of transverse profile with solitary waves

In this section we fit solitary waves to the transverse profiles which are fixed points of the infinite-dimensional map. These fixed point profiles are generated by solving (4.1a) and (4.1b) numerically. First, we describe the saturable case. Typical results are depicted in Fig. 20(a). There we show a transverse intensity profile generated by solving (4.1a) and (4.1b) with parameters set at $F=100$, $l=2$, $\phi=0.4$, $|a(\cdot)|^2=0.008$ (for single soliton). The profile shown is that of the 200th pass, by which time all transients have died out and the profile is certainly a fixed point. In this case, the profile consists of a single pulse on top of an almost flat background. The pulse is accurately represented by a solitary wave as the figure shows. To obtain that fit, we select the amplitude parameter λ to agree with the numerical peak and solve Eq. (5.2) with appropriate boundary conditions. The wings of the actual profile do not approach zero as does the solitary wave; rather, as established in Sec. IV, they approach the lower branch height of the local plane wave hysteresis curve.

Turning to the case of Kerr nonlinearity, it is more difficult to find parameter values for which the map has fixed points. One such fixed point is shown in Fig. 20(b) where we choose $F=0.8$, $|a(0)|^2=0.003$, $l=1.5$, $\phi=0.4$. The dashed curve is the exact soliton solution which has the same amplitude as the observed fixed point. The two curves agree very well except at the wings. As mentioned above, if the input intensity is raised, the wings begin to grow and interact with the central peak. Then the whole profile starts to oscillate. With these parameter values we have not observed multiring fixed points like those obtained in the saturable case; rather, the oscillations seem to persist indefinitely.

The numerical experiments just described show that the central part of the transverse fixed points are well approximated by solitary waves. However, the correct amplitude parameter has been chosen by a numerical fitting procedure. Next, in Secs. V D and VI, we determine the correct amplitude parameter analytically.

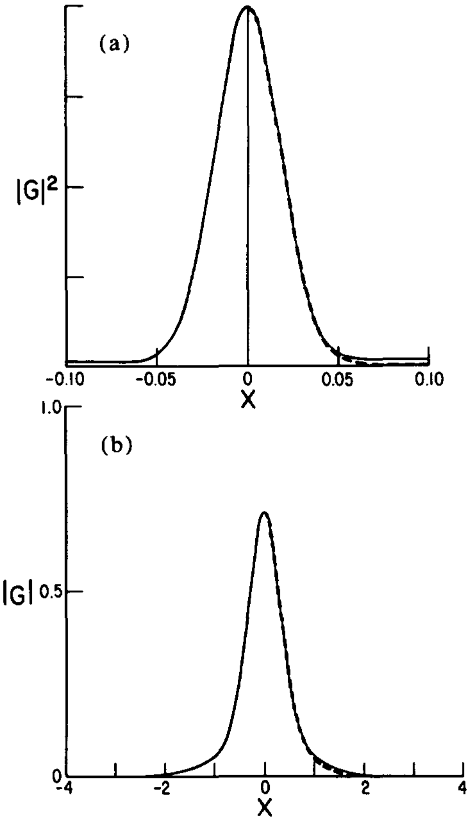

FIG. 20. Comparison of numerically generated single-ring fixed point shapes (solid curves) with solitary waves shapes (dashed curve), of the same amplitude. (a) Saturable nonlinearity [$F=100$, $l=2$, $\phi=0.4$, $|a(\cdot)|^2=0.008$]. (b) Kerr nonlinearity [$F=0.8$, $l=1.5$, $\phi=0.4$, $|a(\cdot)|^2=0.003$].

C. Solitary wave projection formalism

We will use solitary wave perturbation theory to reduce the infinite-dimensional map (4.1a) and (4.1b) to a two-dimensional map. First, some background material. We consider the nonlinear wave equation

$$2i\frac{\partial}{\partial z}G + G_{yy} + N(GG^*)G = 0, \tag{5.7a}$$

a solitary wave solution

$$G_s = S(\lambda y;\lambda)e^{i[(\lambda^2-1)z/2]}, \tag{5.7b}$$

and the linearization of the nonlinear wave equation about the solitary wave ($G = G_s + \tilde{G}$, $|\tilde{G}| \ll 1$),

$$2i\frac{\partial}{\partial z}\tilde{G} + \tilde{G}_{yy} + [N(S^2) + N'(S^2)S^2]\tilde{G}$$
$$+ [N'(S^2)G_s^2]\tilde{G}^* = 0. \tag{5.7c}$$

Symmetries of the nonlinear equation (5.7a) generates solutions of the linear equation (5.7c). In particular, the general solitary wave (5.3) yields four solutions of (5.7c):

$$\tilde{G}^{(1)} = \frac{\partial}{\partial a}G_s\bigg|_{a=0,\ v=0,\ \gamma=0},$$

$$\tilde{G}^{(2)} = \frac{\partial}{\partial \gamma}G_s\bigg|_{a=0,\ v=0,\ \gamma=0},$$

$$\tilde{G}^{(3)} = \frac{\partial}{\partial v} G_s \Big|_{a=0,\quad v=0,\quad \gamma=0}, \tag{5.8}$$

$$\tilde{G}^{(4)} = \frac{\partial}{\partial \lambda} G_s \Big|_{a=0,\quad v=0,\quad \gamma=0}.$$

A basis of generalized solutions of the linear equation (5.7c) exists in which the four solutions (5.8) span a distinguished four-dimensional subspace. In a sense this four-dimensional subspace forms the "soliton component" of the phase space. The details of this decomposition may be found in Appendix B.

D. Solitary wave reduction of the map

Armed with the spectral machinery as described in Appendix B, we return to the infinite-dimensional map (4.1) and consider the possibility of a solitary wave reduction. Given the laser field at the entry point $z = 0$ to the nonlinear medium, we would like to predict which solitary wave emerges at the exit point $z = l$. For this problem to have an answer, the medium must be long enough (l sufficiently large) that the nonlinearity has time to filter the laser field into its asymptotic solitary wave profiles. We restrict our attention to such sufficiently long cavities. Under this restriction a global answer to our problem is known for the Kerr nonlinearity. That is, given any initial data, one can predict, using the inverse scattering transform, exactly which solitons emerge at the end of a Kerr medium. Unfortunately, this mathematical transform method does not apply to more realistic saturable media, and we must content ourselves with a more local problem: For initial data close to a given solitary wave, can we predict which modified solitary wave emerges? This problem takes the mathematical form (for small ϵ)

$$\begin{aligned} &2iG_z + G_{yy} + N(|G|^2)G = 0, \\ &G(y,z=0) = G_s(y,0;\lambda,\gamma,a,v) + \epsilon\tilde{G}_{\text{in}}(y). \end{aligned} \tag{5.9}$$

Which solitary wave emerges? That is, what are the values of its parameters?

In general, these output parameters are not equal to their input values because some of the solitary wave which emerges is hidden in the perturbation $\epsilon\tilde{G}_{\text{in}}$ of the initial data. To see this, assume for the moment that the solitary wave which emerges is equal to the input solitary wave and seek a solution of the form

$$G(y,z) = G_s(y,z;\lambda,\gamma,a,v) + \epsilon\tilde{G}(y,z), \tag{5.10}$$

where to first order in ϵ, $\tilde{G}$ satisfies the linear equation (4.7c). One solution of this equation is

$$\frac{\partial}{\partial\lambda} G_s(y,z;\lambda,\gamma,a,v);$$

thus the $G(y,z)$ can be written as

$$G(y,z;\lambda) = G_s(y,z;\lambda) + c\epsilon\frac{\partial}{\partial\lambda}G_s(y,z;\lambda) + \epsilon\tilde{\tilde{G}},$$

for some constant c which must be determined by the initial data $\tilde{G}_{\text{in}}$. But this equation can be rewritten as

$$G(y,z;\lambda) = G_s(y,z;\lambda^\epsilon = \lambda + \epsilon c) + \epsilon\tilde{\tilde{G}}, \tag{5.11}$$

by a Taylor series expansion of $G_s(\cdot,\cdot;\lambda^\epsilon = \lambda + \epsilon c)$. The solitary wave that emerges has a new parameter $\lambda^\epsilon = \lambda + \epsilon c$, and we must compute its correction c in terms of the initial data $\tilde{G}_{\text{in}}$.

This calculation proceeds as follows: Motivated by (5.11), we replace (5.10) by an ansatz of the form

$$G(y,z) = G_s(y,z;\lambda^\epsilon,\gamma^\epsilon,a^\epsilon,v^\epsilon) + \epsilon\tilde{G}(y,z). \tag{5.12}$$

This amounts to linearizing about the solitary wave which eventually emerges. At this stage in the calculation we do not know the values of the parameters $\lambda^\epsilon,\gamma^\epsilon,a^\epsilon,v^\epsilon$. These must be computed. Now $\tilde{G}$ satisfies the linear problem (5.7c) with initial data

$$\begin{aligned} \tilde{G}(y,z=0) &= \tilde{G}_{\text{in}}(y) + (1/\epsilon)[G_s(y;\lambda,\gamma,a,v) \\ &\qquad - G_s(y;\lambda^\epsilon,\gamma^\epsilon,a^\epsilon,v^\epsilon)] \\ &= \tilde{G}^\epsilon_{\text{in}}(y). \end{aligned} \tag{5.13}$$

We choose the parameters $\lambda^\epsilon,\gamma^\epsilon,a^\epsilon,v^\epsilon$ by demanding that, to first order in ϵ, the $\tilde{G}(y,0)$ contain no solitary wave. This is accomplished by demanding the $\tilde{G}(y,0)$ be orthogonal in an appropriate sense (see Appendix B) to the four solutions (5.8). Explicitly, we write

$$\tilde{G}(y,z) = g(y,z)e^{i[vy + [(\lambda^2 - 1 - v^2)/2]z + \gamma]},$$

$$g = U + iV, \quad \mathbf{g} = \begin{pmatrix} U \\ V \end{pmatrix},$$

and demand

$$\begin{aligned} &(S^\epsilon, U(\cdot,0)) = (S^\epsilon, \operatorname{Re} g(\cdot,z=0)) = 0, \\ &(yS^\epsilon, U(\cdot,0)) = (yS^\epsilon, \operatorname{Re} g(\cdot,z=0)) = 0, \\ &(S^\epsilon_y, V(\cdot,0)) = (S^\epsilon_y, \operatorname{Im} g(\cdot,z=0)) = 0, \\ &(yS^\epsilon_y + \lambda^\epsilon S^\epsilon_\lambda, V(\cdot,0)) \\ &\qquad = (yS^\epsilon_y + \lambda^\epsilon S^\epsilon_\lambda, \operatorname{Im} g(\cdot,z=0)) = 0, \end{aligned} \tag{5.14}$$

where

$$g(y,0) = e^{-i[vy+\gamma]}\tilde{G}^\epsilon_{\text{in}}(y).$$

We now apply criteria (5.14) to the map for the optical bistability problem:

$$\begin{aligned} &2i\frac{\partial}{\partial z}G_{n+1} + G_{n+1,yy} + N(|G_{n+1}|^2)G_{n+1} = 0, \\ &G_{n+1}(y,0) = a(y) + Re^{i\phi}e^{i[[(\lambda_n^2 - 1)/2]l + \gamma_n]}S(\lambda_n y;\lambda_n). \end{aligned} \tag{5.15}$$

In writing (5.15), we have assumed that the output of the nth pass down the nonlinear medium is a pure solitary wave. All other modes, such as radiation, have been neglected. In addition, we have elected to examine an output which is symmetric about the $y = 0$ axis. Thus the parameters a and v may be ignored by symmetry considerations. Since $R \simeq 1$ and $a \ll 1$, initial data (5.15b) may be treated as a small perturbation of a solitary wave. (Here the phase factor $[\phi + (\lambda_n^2 - 1)l/2 + \gamma_n]$ merely changes the phase in the solitary wave.) We use (5.14) to predict values for the parameters of the solitary wave that emerges after the next pass:

$$\begin{aligned} (S_{n+1},S_{n+1}) = {}&(A_{n+1},S_{n+1})\cos\gamma_{n+1} \\ &+ R\cos\Gamma_{n,n+1}(S_{n+1},S_{n,n+1}), \end{aligned}$$

$$
\begin{aligned}
0 = & -(A_{n+1},\rho_{n+1})\sin\gamma_{n+1} \\
& + R\sin\Gamma_{n,n+1}(\rho_{n+1},S_{n,n+1}),
\end{aligned} \tag{5.16}
$$

where

$$
\begin{aligned}
A_{n+1} &= a(\theta/\lambda_{n+1}), \\
S_{n+1} &= S(\theta;\lambda_{n+1}), \\
\Gamma_{n,n+1} &= \phi + (\gamma_n - \gamma_{n+1}) + (l/2)(\lambda_n^2 - 1), \\
S_{n,n+1} &= S(\lambda_n\theta/\lambda_{n+1};\lambda_n), \\
\rho_{n+1} &= \frac{1}{\lambda_{n+1}}\theta S_\theta(\theta;\lambda_{n+1}) + \frac{\partial}{\partial\lambda}S(\theta,\lambda)\Big|_{\lambda=\lambda_{n+1}}.
\end{aligned} \tag{5.17}
$$

Equation (5.16) defines a two-dimensional real map on the solitary waves parameters (λ,γ):

$$(\lambda_n,\gamma_n)\to(\lambda_{n+1},\gamma_{n+1}). \tag{5.18}$$

We have used solitary wave perturbation theory to reduce the infinite-dimensional map (3.1) to a two-dimensional one (5.17).

The equation for the fixed points (λ,γ) is much simpler than the map itself,

$$
\begin{aligned}
(S,S) &= (A,S)\cos\gamma + R\cos\Gamma(S,S), \\
0 &= -(A,\rho)\sin\gamma + R\sin\Gamma(\rho,S),
\end{aligned} \tag{5.19}
$$

where

$$S = S(\theta;\lambda), \quad A = a(\theta/\lambda),$$

$$\Gamma = \phi + (l/2)(\lambda^2 - 1), \quad \rho = \frac{1}{\lambda}\theta S_\theta + \frac{\partial}{\partial\lambda}S(\theta;\lambda).$$

VI. REDUCED MAP ON THE PARAMETERS OF THE SOLITARY WAVE

The main result of Sec. V is the map

$$(\lambda_n,\gamma_n)\to(\lambda_{n+1},\gamma_{n+1}) \tag{6.1}$$

on the amplitude and phase parameters, which is given explicitly by (5.16). The fixed points of this map satisfy (5.19); thus (5.19) predicts the parameter values for the solitary wave that finally emerges after many passes through the nonlinear medium.

A. Reduced map—saturable case

In general Eq. (5.19) for the fixed points is quite implicit. First, we solve it numerically for the saturable case. Typical results are shown in Fig. 21. We emphasize that there are no free parameters in this theory. The theory rigidly predicts the amplitude of the solitary wave that emerges. The results (Fig. 21) are very accurate.

B. Reduced map—Kerr case

In the Kerr case, much more can be done analytically, primarily because of the explicit formula for the solitary wave,

$$S(\theta;\lambda) = \lambda\,\text{sech}\,\theta. \tag{6.2}$$

Using this formula we place the reduced map (5.16) in the form

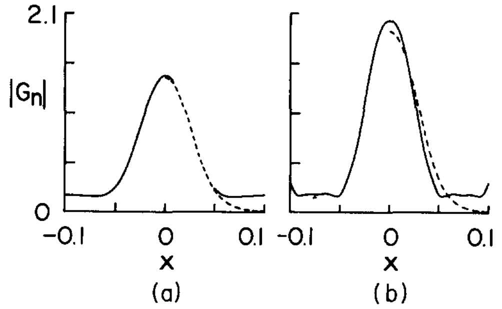

FIG. 21. Comparison of solitary wave fixed points (dashed curves) of the reduced map [Eq. (5.19)] with the numerically generated shapes for a saturable nonlinearity. (a) Comparison with the single-ring shape (solid curve) [$F = 200$, $l = 2$, $\phi = 0.4$, $a(0) = 0.1$]. (b) Comparison with the central ring of the seven-ring fixed points [$F = 200$, $l = 2$, $\phi = 0.4$, $a(\cdot) = 0.19$] (see Ref. 3).

$$
\begin{aligned}
\lambda_{n+1} = & A_s(\lambda_{n+1})\cos\gamma_{n+1} \\
& + RB_s(\lambda_n/\lambda_{n+1})\cos(\Gamma_{n,n+1})\lambda_n, \\
0 = & -A_p(\lambda_{n+1})\sin\gamma_{n+1} \\
& + RB_p(\lambda_n/\lambda_{n+1})\sin(\Gamma_{n,n+1})\lambda_n,
\end{aligned} \tag{6.3}
$$

where

$$
\begin{aligned}
A_s(\lambda_{n+1}) &= \frac{1}{2}\int \text{sech}\,\theta a\left(\frac{\theta}{\lambda_{n+1}}\right)d\theta, \\
A_p(\lambda_n) &= \int(\theta\,\text{sech}\,\theta)_\theta a\left(\frac{\theta}{\lambda_{n+1}}\right)d\theta, \\
B_s\left(\frac{\lambda_n}{\lambda_{n+1}}\right) &= \frac{1}{2}\int \text{sech}\,\theta\,\text{sech}\left(\frac{\lambda_n\theta}{\lambda_{n+1}}\right)d\theta, \\
B_p\left(\frac{\lambda_n}{\lambda_{n+1}}\right) &= \int(\theta\,\text{sech}\,\theta)_\theta\,\text{sech}\left(\frac{\lambda_n\theta}{\lambda_{n+1}}\right)d\theta, \\
\Gamma_{n,n+1} &= \phi + (\gamma_n - \gamma_{n+1}) + (l/2)(\lambda_n^2 - 1).
\end{aligned}
$$

Map (6.3) should be compared with the plane wave map for the Kerr nonlinearity:

$$g_{n+1} = a + Re^{i[\phi + (l/2)(2|g_n|^2 - 1)]}g_n, \tag{6.4}$$

which may be rewritten in the form ($g = \lambda e^{i\gamma}$)

$$
\begin{aligned}
\lambda_{n+1} &= a\cos\gamma_{n+1} + R\cos(\Gamma_{n,n+1})\lambda_n, \\
0 &= -a\sin\gamma_{n+1} + R\sin(\Gamma_{n,n+1})\lambda_n.
\end{aligned} \tag{6.4'}
$$

In this Kerr case, the map on solitary wave parameters,

$$
\begin{aligned}
\lambda_{n+1} = & A_s(\lambda_{n+1})\cos\gamma_{n+1} \\
& + RB_s(\lambda_n/\lambda_{n+1})\cos(\Gamma_{n,n+1})\lambda_n, \\
0 = & -A_p(\lambda_{n+1})\sin\gamma_{n+1} \\
& + RB_p(\lambda_n/\lambda_{n+1})\sin(\Gamma_{n,n+1})\lambda_n,
\end{aligned} \tag{6.3}
$$

and the plane wave are very similar. The main difference is that constants in the plane wave case are replaced by projections over solitary wave profiles. These projections make the map (6.3) slightly more implicit than its plane wave counterpart because of the dependence on λ_{n+1} on the right-hand side. More importantly, a symmetry in the plane wave map [$a\cos\gamma_{n+1}$, $-a\sin\gamma_{n+1}$] is drastically broken by these projections. To see this, realize that the solitary waves which evolve are typically narrow when compared to the input Gaussian. In this case, the projections of the Gaussian $a(y)$ can be estimated:

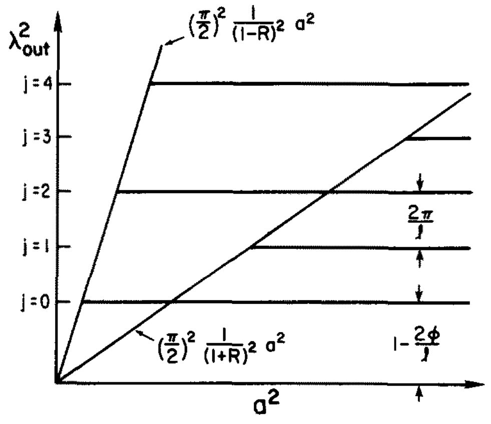

FIG. 22. Fixed points of the reduced map for the "constant Kerr" case [Eq. (6.6)].

$$A_s(\lambda) = \frac{1}{2}\int \operatorname{sech}\theta\, a\Big(\frac{\theta}{\lambda}\Big)d\theta \simeq \frac{\pi}{2}a(0),$$
$$A_\rho(\lambda) \simeq (\theta \operatorname{sech}\theta)_\theta a(\theta/\lambda)d\theta \simeq 0. \qquad (6.5)$$

We use this calculation to introduce a third map, which we call the "contant-Kerr" case $[a = a(0)]$:

$$\lambda_{n+1} = (\pi/2)a\cos\gamma_{n+1} + RB_s(\lambda_n/\lambda_{n+1})\cos(\Gamma_{n,n+1})\lambda_n, \qquad (6.6a)$$
$$0 = +RB_\rho(\lambda_n/\lambda_{n+1})\sin(\Gamma_{n,n+1})\lambda_n.$$

This last map is rather easy to analyze. Its fixed points (λ,γ) satisfy

$$\lambda = (\pi/2)a\cos\gamma + R(\cos\Gamma)\lambda, \qquad (6.6b)$$
$$0 = (\sin\Gamma)\lambda, \quad \Gamma = \phi + (l/2)(\lambda^2 - 1),$$

which can be solved explicitly to yield

$$\lambda = 0, \quad \Gamma = j\pi \Rightarrow \lambda_j = \sqrt{1 + 2(j\pi - \phi)/l}, \qquad (6.6c)$$
$$\cos\gamma_j = (2/\pi a)[1 - (-1)^j R\,]\lambda_j.$$

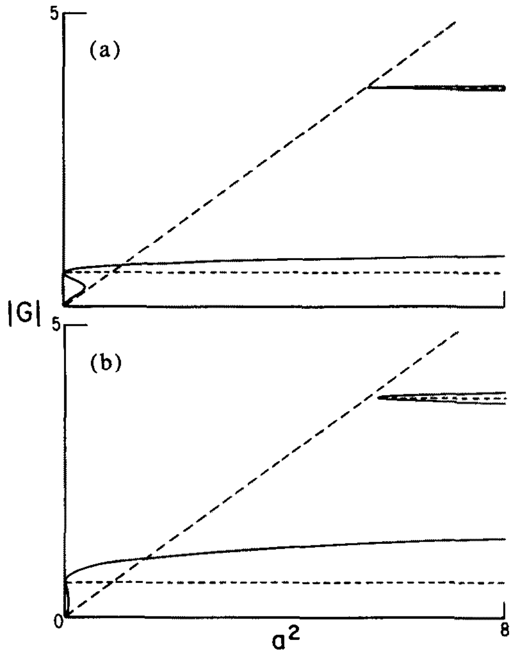

FIG. 23. Fixed points of the reduced Kerr map [Eq. (6.3)] showing the breaking of degeneracy by the transverse spatial dependence of $a(\cdot)$. (a) $F=5$, $\phi=0.4$, $l=2.0$, $R=0.9$. (b) $F=1$, $\phi=0.4$, $l=2$, $R=0.9$.

These are sketched in Fig. 22. Notice that the only a^2 dependence is a lower cutoff which guarantees $|\cos\gamma_j| < 1$. These cutoffs are the maximum–minimum responses of the plane wave map, Eq. (3.7). Notice also that each curve, except $\lambda = 0$, stands for two fixed points $(\lambda_j, \pm\gamma_j)$. As approximation (6.5) is removed, this degeneracy is broken and the curves develop a dependence on the amplitude a^2. These are pictured in Fig. 23. Next, we linearize this map about a fixed point (λ_j,γ_j):

$$\begin{pmatrix}\tilde\lambda_{n+1}\\ \tilde\gamma_{n+1}\end{pmatrix} = T(\lambda,\gamma)\begin{pmatrix}\tilde\lambda_n\\ \tilde\gamma_n\end{pmatrix}, \qquad (6.6d)$$

where

$$\lambda = \lambda_j \neq 0, \quad \gamma = \gamma_j,$$
$$T(\lambda,\gamma) = \begin{bmatrix} \dfrac{-l(\pi a/2)\lambda\sin\gamma + (-1)^j R/2}{(1 + (-1)^{j+1}R/2)} & \dfrac{-(\pi a/2)\sin\gamma}{(1+(-1)^{j+1}R/2)} \\ l\lambda & 1 \end{bmatrix}. \qquad (6.6e)$$

Now the Jacobian at this fixed point is given by

$$\det T(\lambda,\gamma) = (-1)^j R/2/[1 - (-1)^j R/2]. \qquad (6.6f)$$

This is approximately equal to R^2 if j is even and far from R^2 if j is odd. For this reason, we restrict our attention to the even values of j. For even j, the eigenvalues of T satisfy

$$\mu^2 - \frac{1 - (l\pi a/2)\lambda\sin\gamma}{1 - R/2}\mu + \frac{R/2}{1 - R/2} = 0. \qquad (6.6g)$$

As in the case of the plane wave map, these eigenvalues are either both real or are conjugates of each other. In the latter case, they always satisfy $|\mu| < 1$, and hence the corresponding fixed points are stable. When the eigenvalues are real, they satisfy $\mu_1\mu_2 = R/(2 - R)$. They both can be less than 1 or one can pass through either $+1$ or -1. The latter case indicates a period doubling instability of the fixed point.

Consider a fixed point (λ_j,γ_j) and its sister $(\lambda_j, -\gamma_j)$. We call $(\lambda_j,\gamma_j > 0)$ the "lower jth branch" and $(\lambda_j, -\gamma_j)$ the "upper jth branch" because of their locations once the degeneracy is split by removing (6.5). One can show that the lower branch is always unstable. The upper branch has the stability depicted in Fig. 24. A curve can be drawn, to the immediate right of which a period-2 bifurcation occurs (Fig. 24). Some parameter values at which the period-2 bifurcation occurs are listed in Table I.

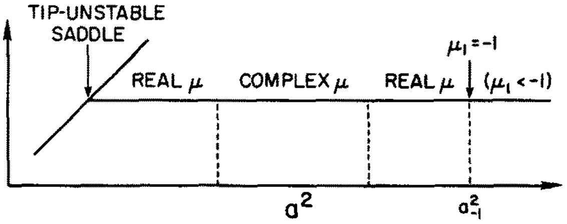

FIG. 24. Stability of the upper branch of the eigenvalue pair as a function of increasing pump intensity $a(\cdot)^2$ for the "constant" Kerr map. The eigenvalue pair is denoted by μ along the horizontal line. The period doubling bifurcation occurs at a^2_{-1} $(\mu_1=-1)$.

When a is nonconstant, the approximation (6.5) applies no more and the fixed point of the map (6.3) has to be evaluated numerically. To be consistent with (2.2a) and (2.2b), we choose $a(y)=a(0)\exp[-[\ln 2/4\pi Fl]y^2]$. A result of such computation is shown in Fig. 25 where the solitary wave fixed point is plotted as a dashed line and, for comparison, the plane wave fixed point is also plotted in a solid line. The parameter values are $F=0.8$, $l=1.5$, $\phi=0.4$.

Keeping the same parameter values, we solved the equation numerically and compared resulting fixed points with analytical ones. Two such attempts are shown in Figs. 26 and 27, where dashed curves represent predicted solitary wave fixed points. In the first case, the relative error in amplitude is about 2.9%; in the second case, it is about 6.7%.

The discrepancies come from two sources. One is that the observed fixed points are not pure solitons but rather a combination of a single soliton and a lower branch fixed point in the wings, where the latter presumably "pushes up" the former resulting in a larger amplitude than the prediction. The other is more subtle. In deriving the map we assumed the form (5.10), where at each pass the boundary condition is updated by a soliton plus a small perturbation. Since the explicit expression of this perturbation term is known in our formalism, we can compute its size relative to the soliton. The computation shows that for certain parameter values its size becomes large. The region where the perturbation stays small turns out to be confined to the leftmost part of the hystersis curves, which is shown shaded in Fig. 28. The crosses in the same figure represent the observed fixed points. At the "tip" of the curve the relative size of the perturbation becomes minimum, where we expect the best fit.

TABLE I. Parameter values for period-2 bifurcations.

$\phi=0.4,\quad l=2\pi,\quad R=0.9$	
$j=0$	$a^2_{-1}=0.0506$
$j=2$	$a^2_{-1}=0.0260$
$j=4$	$a^2_{-1}=0.0282$
$j=6$	$a^2_{-1}=0.0338$
$j=8$	$a^2_{-1}=0.0406$
$\phi=0.4,\quad l=2,\quad R=0.9$	
$j=0$	$a^2_{-1}=0.677$
$j=2$	$a^2_{-1}=0.0867$
$j=4$	$a^2_{-1}=0.0841$
$j=6$	$a^2_{-1}=0.0994$

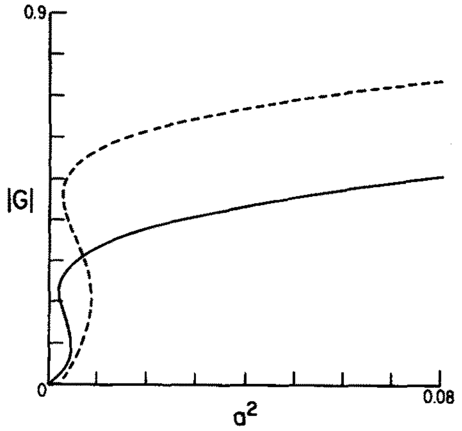

FIG. 25. Comparison of the numerically generated fixed point curve of the reduced Kerr map (dashed) with the corresponding fixed point curve of plane wave map (solid). Parameters: $F=0.8$ (dashed curve), $l=1.5$, $\phi=0.4$, and $R=0.9$.

Numerical experiments such as these establish that, when solitary wave fixed points occur, their amplitudes (and widths) are accurately predicted by the reduced maps, (5.19) in the saturable case and (6.3) in the Kerr case. However, the experiments also show that the stability of these fixed points is not accurately captured by these reduced maps. Accurate stability calculations require the inclusion of more degrees of freedom than just the (two-parameter) solitary wave ansatz. Stability calculations are discussed in Ref. 4.

VII. ANALYTICAL FORMULA FOR SOLITON PLUS FLAT BACKGROUND

In this section we derive an analytical solution of the nonlinear Schrödinger equation with Kerr nonlinearity which consists of a soliton plus a flat background. This solution has enough freedom to fit both a central peak and wings of the fixed point profile. For simplicity, consider

$$iq_t+q_{xx}+2|q|^2q=0, \tag{7.1}$$

which can be easily transformed into (2.2a). We apply a

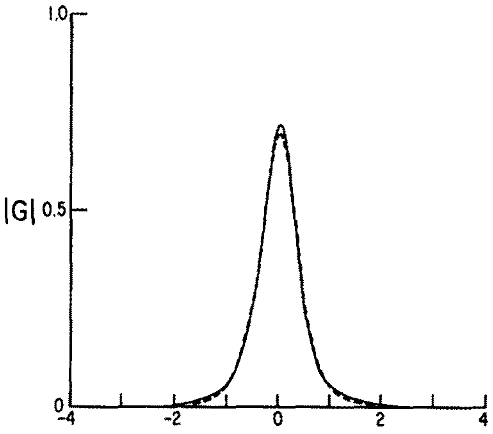

FIG. 26. Comparison of numerically generated fixed points of the infinite-dimensional map (solid curve) with the corresponding fixed point of the reduced Kerr map (dashed curve). Same parameters are used as in Fig. 25 with $a(\cdot)^2=0.003$.

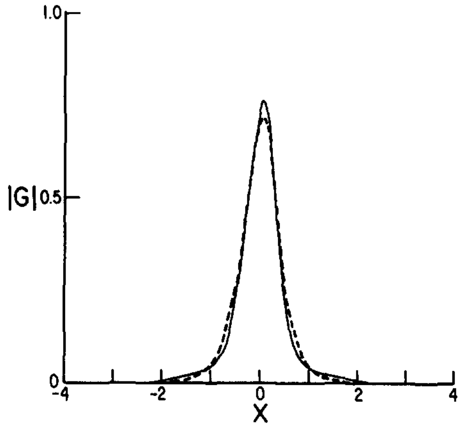

FIG. 27. Similar fixed point comparison as in Fig. 26 except that $a(\cdot)^2 = 0.004$. The increased discrepancy between fixed points is evident with increasing $a(\cdot)^2$.

Bäcklund transformation from the flat background solution $q_0 = Ke^{i2|K|^2t}$ of (7.1) to create the desired formula.[11]

The Lax equations in this case reduce to the following pair of Riccati equations:

$$\alpha_x = q_0^* \alpha^2 - 2i\zeta\alpha + q_0, \tag{7.2a}$$

$$\alpha_t = -C\alpha^2 + 2A\alpha + B, \tag{7.2b}$$

where

$$A = -2i\zeta^2 + iq_0q_0^*, \quad B = 2q_0\zeta + iq_{0,x},$$
$$C = -2q_0^*\zeta + iq_{0,x}^*, \quad \zeta = \xi + i\eta. \tag{7.3}$$

Then the new solution q of (7.1) becomes

$$q = q_0 + 4\eta\alpha/(1 + |\alpha|^2). \tag{7.4a}$$

Solving (7.2a) and (7.2b), we get

$$q = K + \frac{4\eta(\bar{f} + B|f|^2)}{|f|^2 + |1 + Bf|^2} e^{2i|K|^2t}, \tag{7.4b}$$

where

$$f = K^*/2iw + e^{2iw(x + \zeta t) + \Omega},$$
$$B = i(\zeta - w)/K^*,$$
$$w = (\zeta^2 + |K|^2)^{1/2},$$
$$\Omega = \text{const}.$$

We can show (1) as $K \to 0$, q reduces to a single soliton formula; (2) for the special case where $\zeta = i\eta$ and $\eta^2 > |K|^2$, $\lim_{x \to \pm\infty} |q| = |K|$. Also, we have checked, by direct substitution, that is indeed a solution. The periodic behavior of this solution is shown below.

The characteristic oscillation of the center and wings seen in Fig. 29 has been observed in actual fixed points as they propagate through an extended medium. Numerical fits have been difficult and we are unable to present them here. Nevertheless, formula (7.4) provides us with an analytical solution which has enough freedom to describe the complete profile of the single solitary wave fixed point, including both its central peak and its wings.

VIII. CONCLUSION

In this first of a series of papers, we have studied the dynamics of an electromagnetic field in an optical ring cavity. In the present article our studies have been restricted to one transverse spatial dimension. Specifically, we have identified numerically generated fixed points of an infinite-dimensional map (which models the physical problem) with fixed points of reduced maps in the solitary wave (soliton) parameters, for saturable (Kerr) nonlinearities. A math-

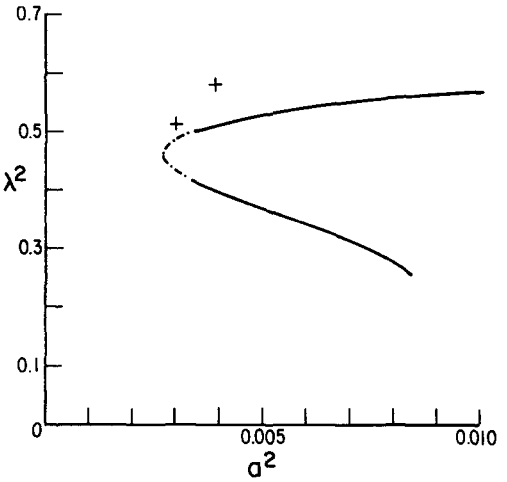

FIG. 28. Upper branch soliton fixed point curve of the reduced Kerr map. The dot–dashed part denotes the region over which the perturbation is small (see text). The two crosses denote the numerically generated peak intensities. ($\lambda^2 = G^2$) from the infinite-dimensional Kerr map.

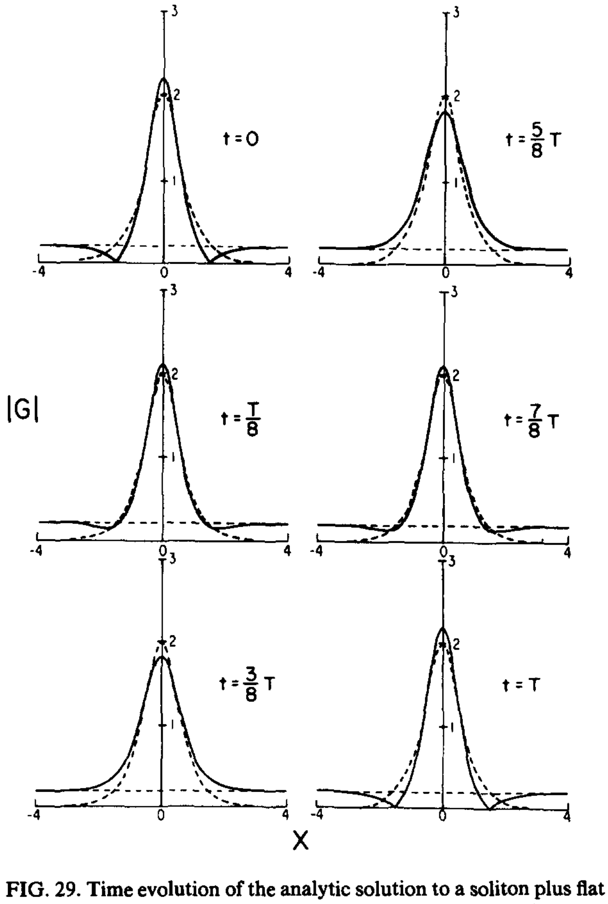

FIG. 29. Time evolution of the analytic solution to a soliton plus flat background (dashed curves) over a single period T of oscillation. The solid curves represent the superposition of both pieces.

ematical projection formalism has been developed to derive the reduced maps, the fixed points of which accurately predict the output field shapes for the saturable case and give good agreement over a more restrictive parameter range in the Kerr case. In the saturable case there is little freedom in the natural state of propagation which seems to produce, in the full map, spatially isolated entities which phase lock. In the Kerr case there are many independent N-soliton states of propagation which complicate the response in the full map. This rich variety of states makes the projection onto a single solitary wave much less robust in this latter case.

Stability of these fixed points will be addressed in a future publication. As a fixed point loses stability because of an increased stress, the output experiences a fascinating transition into a modulational chaos. Correlation of this temporal chaos with additional transverse spatial structure is currently under investigation. Preliminary results may be found in Ref. 2.

The situation in two (2) transverse dimensions is even more interesting. Coherent spatial structures again play a central role, but the chaotic state is much more severe. Preliminary results are summarized in Ref. 2.

APPENDIX A: PHYSICAL AND MATHEMATICAL DESCRIPTION OF THE DYNAMICS OF RING CAVITIES

In this appendix, we provide, for the readers' convenience a derivation from the first principles of the model equation. Much of this material can be found in texts on quantum optics, each with their own notation.

The geometry of the problem is a ring cavity in which the signal always propagates in one direction only. The dynamics of the electromagnetic field and the nonlinear medium are described by the Maxwell–Bloch equations.

If the displacement field $\mathbf{D}$ is written as $\mathbf{E}+4\pi\mathbf{P}$ (Gaussian units are used throughout), where $\mathbf{E}$ is the electric field vector and $\mathbf{P}$ the polarization induced by $\mathbf{E}$, then $\mathbf{E}$ satisfies

$$\nabla^2\mathbf{E}-\frac{1}{c^2}\frac{\partial^2\mathbf{E}}{\partial t^2}=\frac{4\pi}{c^2}\frac{\partial^2\mathbf{P}}{\partial t^2}-4\pi\nabla(\nabla\cdot\mathbf{P}). \tag{A1}$$

The polarization $\mathbf{P}$ induced by the electric field is caused by the excitation of the atoms in the medium into a higher energy state. We assume a two-level medium in which the electron can occupy one of two states, a or b. (Spontaneous emission from the higher energy state is important and account will be taken of this type of loss). The electron wave function is written

$$\Psi(\mathbf{r},\mathbf{R},t)=a(\mathbf{r},t)\psi_a(\mathbf{R})+b(\mathbf{r},t)\psi_b(\mathbf{R}), \tag{A2}$$

where the eigenstates ψ_a,ψ_b are normalized such that

$$\int\psi_a^*\psi_b\,d\mathbf{R}=\delta_{ab},\quad e\int\mathbf{R}\psi_a^*\psi_b\,d\mathbf{R}=e\int\mathbf{R}\psi_a\psi_b\,d\mathbf{R}=\mathbf{p}. \tag{A3a}$$

Symmetry of the states ψ_a,ψ_b implies

$$\int\mathbf{R}\psi_a\psi_a^*\,d\mathbf{R}=\int\mathbf{R}\psi_b\psi_b^*\,d\mathbf{R}=0. \tag{A3b}$$

The positive vectors $\mathbf{r}$ and $\mathbf{R}$ refer to the center of the atom and the relative position of the electron, respectively (see Fig. 30). The polarization per atom is

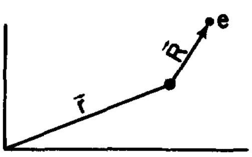

FIG. 30. An atom in the laboratory frame of reference: nucleus at $\mathbf{r}$ and electron at $\mathbf{r}+\mathbf{R}$.

$$e\int\mathbf{R}\Psi\Psi^*\,d\mathbf{R}=\mathbf{p}(ab^*+a^*b). \tag{A4}$$

The wave function Ψ satisfies the Heisenberg equation

$$ih\frac{\partial\Psi}{\partial t}=(H_0+U)\Psi, \tag{A5}$$

where the Hamiltonian $H=H_0+U$ consists of an unperturbed component H_0,

$$H_0\psi_a=h\omega_a\psi_a,\quad H_0\psi_b=h\omega_b\psi_b, \tag{A6}$$

and a potential U induced by the field $\mathbf{E}(\mathbf{r},t)$,

$$U=-e\mathbf{E}\cdot\mathbf{R}. \tag{A7}$$

From (A5) and (A2), the probability amplitudes satisfy the equations

$$a_t=-i\omega_a a+[(i\mathbf{E}\cdot\mathbf{p})/h]b, \tag{A8a}$$

$$b_t=-i\omega_b b+[(i\mathbf{E}\cdot\mathbf{p})/h]a. \tag{A8b}$$

We will delay the introduction of the losses due to spontaneous emission until the subsection entitled homogeneous broadening, but we stress that the situation studied in this paper is dominated by their relaxation effects.

Our first goal is to develop equations for the quadratic quantities

$$Q_r=ab^*+a^*b,\quad Q_i=i(ab^*-a^*b),\qquad \eta=aa^*-bb^*, \tag{A9}$$

which measure polarization, its conjugate, and the population inversion of the system. If there is no net loss from the a, b states, then the probability of finding the atom in one of the two states is unity,

$$aa^*+bb^*=1. \tag{A10}$$

A value of $\eta=-1$ means that all the atoms are in the lower state b.

In order to write equations for Q_r, Q_i, and η it is convenient to remove the ultrafast time scale $((\omega_a+\omega_b)/2)^{-1}$ from the dynamics of the probability amplitudes a and b by setting

$$\begin{pmatrix}a\\b\end{pmatrix}=e^{-i[(\omega_a+\omega_b)/2]t}u, \tag{A11}$$

whereupon $u=(u_1,u_2)^T$ satisfies

$$u_t=\begin{pmatrix}-i\omega_{ab}/2 & i(\mathbf{E}\cdot\mathbf{p})/h\\ -i(\mathbf{E}\cdot\mathbf{p})/h & i\omega_{ab}/2\end{pmatrix}u \tag{A12}$$

where $\omega_{ab}=\omega_a-\omega_a$ is of the same order of magnitude ($\sim10^{15}$ sec^{-1}, corresponding to a wavelength of 6000 Å, close to the D_1, D_2 lines of sodium) as ω is the carrier frequency of the electric field $\mathbf{E}(\mathbf{r},t)$. For situations discussed in the paper, the electric field magnitude is such that the Rabi frequency $(2p/h)E$ is approximately the same order of magnitude as the detuning $(\omega_{ab} - \omega) \sim 10^8\text{–}10^{11}\ \text{sec}^{-1}$ of the system. Therefore the ratio $\rho = 2pE/h\omega_{ab}$ of the off-diagonal to diagonal terms in (A12) is $\sim 10^{-7} \to 10^{-4}$ and small. Consequently it is natural to solve (A12) iteratively by setting

$$u = u^{(0)} + \rho u^{(1)} + \rho^2 u^{(2)} + \cdots, \tag{A13}$$

with

$$u^{(0)} = \begin{pmatrix} e^{-i\omega t/2} & 0 \\ 0 & e^{i\omega t/2} \end{pmatrix} U(t), \tag{A14}$$

where

$$\mathbf{E}(r,t) = \mathbf{F}(\mathbf{r},t)e^{-i\omega t} + \mathbf{F}^*(\mathbf{r},t)e^{i\omega t} \tag{A15}$$

and U, F, and its complex conjugate F^* vary slowly over times of order 10^{-15} sec. It is not hard to see that the slow dependence of U on t must be chosen to be

$$U_t = \begin{pmatrix} -i\zeta & (i\mathbf{F}\cdot\mathbf{p})/h \\ (i\mathbf{F}^*\cdot\mathbf{p})/h & i\zeta \end{pmatrix} U, \quad 2\zeta = \omega_{ab} - \omega, \tag{A16}$$

in order that $u^{(1)}$ contains no secular terms proportional to t. This is called the "rotating wave" approximation in the physics literature. The reason for the potential appearance of secular terms is the near resonance between the frequency ω of the applied field and the two-level frequency ω_{ab}, a resonance which would cause the asymptotic expansion (A13) to cease to be valid after times $t \sim (\omega_{ab} - \omega)^{-1}$. The second harmonic terms $e^{\pm 2i\omega t}$ do not cause any such problems and can be included in $u^{(1)}$.

From (A9), (A11), (A14), and (A16) we note that

$$Q_r = U_1 U_2^* e^{-i\omega t} + U_1^* U_2 e^{i\omega t}, \tag{A17a}$$

$$Q_i = iU_1 U_2^* e^{-i\omega t} - iU_1^* U_2 e^{i\omega t}, \tag{A17b}$$

$$\eta = U_1 U_1^* - U_2 U_2^*, \tag{A17c}$$

and

$$T = aa^* + bb^* = U_1 U_1^* + U_2 U_2^*. \tag{A17d}$$

The neglected terms in (A17) are of the order $(\omega_{ab} - \omega)/\omega_{ab}$. It is convenient to write equations for the complex polarization

$$\Lambda = 2U_1 U_2^* \tag{A18}$$

rather than for Q_r, Q_i separately. We find

$$\Lambda_t = -2i\zeta\Lambda - 2i[(\mathbf{F}\cdot\mathbf{p})/h]\eta, \tag{A19}$$

$$\eta_t = (i\mathbf{p}/h)\cdot(\mathbf{F}\Lambda^* - \mathbf{F}^*\Lambda), \tag{A20}$$

$$T_t = 0. \tag{A21}$$

Equations (A19)–(A21) are called the Bloch equations. They allow us to compute the induced polarization per individual atom

$$\mathbf{P}_A = \tfrac{1}{2}\mathbf{p}(\Lambda e^{-i\omega t} + \Lambda^* e^{+i\omega t}) \tag{A22}$$

as a functional of the applied electric field $\mathbf{E}$,

$$\mathbf{E} = \mathbf{F}e^{-i\omega t} + \mathbf{F}^* e^{i\omega t}. \tag{A23}$$

The evolution of the electric field is given by (A1). The polarization $\mathbf{P}$ in (A1) is the sum over the polarizations of all the individual atoms and we will write this as

$$\mathbf{P} = \langle \mathbf{P}_A \rangle. \tag{A24}$$

Before explaining how the sum $\langle\ \rangle$ is taken, let us make two straightforward observations. The first is that the induced polarization $\mathbf{P}$ is parallel to the applied field and therefore we can write $\mathbf{E}$ as $E\hat{e}$, $\mathbf{p}$ as $p\hat{e}$, where $\hat{e}$ is a unit vector perpendicular to the direction of propagation, and (A1) becomes a scalar equation. Since P_A has the form (A22) and Λ varies on the time scale $(\omega_{ab} - \omega)^{-1}$ we can replace P_{tt} in (A1) by $-\omega^2 P$. Also since the variation of both E and P (which are perpendicular to the direction of propagation) in the direction of propagation is small, we can neglect the divergence term on the right-hand side of Eq. (A1) which now becomes

$$\nabla^2 E - \frac{1}{c^2}\frac{\partial^2 E}{\partial t^2} = -\frac{4\pi}{c^2}\omega^2 P. \tag{A25}$$

Equations (A19)–(A21), (A22), (A24), and (A25) form a closed set of equations for the electric field, induced polarization, and population inversion of the system.

1. The phenomenon of inhomogeneous broadening

We now discuss how to compute the collective polarization P. If every atom behaved in exactly the same way, Eq. (A24) would simply read

$$P = nP_A, \tag{A26}$$

where n is the density (number per cm^3) of atoms. However, because of the random motion of the atoms, each sees the frequency of the applied field Doppler shifted by an amount $\mathbf{k}\cdot\mathbf{v}_A$ (where $k = 2\pi/\lambda$ is the wave number of the carrier wave and $\mathbf{v}_A$ is the velocity of atom A). Equivalently this means that the effective two-level frequency is distributed over a set of frequencies $\omega_{ab}^{(0)} + \mathbf{k}\cdot\mathbf{u}$, where $h\omega_{ab}^{(0)}$ is the difference in energy levels a and b for an atom at rest. Therefore the frequency difference parameter

$$2\zeta = \omega_{ab} - \omega$$

is distributed over a range of frequencies, characterized by a probability distribution $g(2\zeta)$ reflecting the distribution of velocities of the atoms. Given $\int g(2\zeta)d(2\zeta) = 1$, we now calculate (A24) as

$$P(\mathbf{r},t) = n\int g(2\zeta)P_A(\zeta,\mathbf{r},t)d(2\zeta). \tag{A27}$$

Frequently the distribution $g(2\zeta)$ is approximated by the Lorentzian

$$g(2\zeta) = (\Gamma/\pi)\{1/[\Gamma^2 + 4(\zeta - \zeta_0)^2]\}, \tag{A28}$$

where Γ is the line width and $2\zeta_0$ measures the distance of the center of the distribution from the frequency of exact resonance. We call the situation in which the Lorentzian linewidth Γ approaches zero the *sharp line limit* and in this case the distribution function $g(2\zeta)$ becomes the Dirac delta function $\delta(2(\zeta - \zeta_0))$.

2. The sharp line, on resonance limit

We now calculate what happens in this limit ($\Gamma \to 0$) when the distribution is centered about the exact resonance frequency, namely where $\zeta_0 = 0$. In this case

$$\Lambda_t = -2i(Fp/h)\eta, \tag{A29}$$

$$\eta_t = (ip/h)(F\Lambda^* - F^*\Lambda), \tag{A30}$$

and

$$\nabla^2 E - (1/c^2)E_{tt}$$
$$= -(2\pi\omega^2/c^2)np(\Lambda e^{-i\omega t} + \Lambda^* e^{+i\omega t}). \tag{A31}$$

Let us now write

$$E(\mathbf{r},t) = \epsilon(r,t)\sin(\omega t - kz), \tag{A32}$$

where $\epsilon(\mathbf{r},t)$, the electric field envelope, varies slowly with respect to the carrier wave in the propagation direction z and in time t, and not at all in the transverse directions x and y; i.e.,

$$\frac{\partial\epsilon}{\partial t} \ll \omega\epsilon, \quad \frac{\partial\epsilon}{\partial z} \ll k\epsilon.$$

We obtain, on comparing the coefficients of $\cos(\omega t - kz)$,

$$\epsilon_z + (1/c)\epsilon_t = (2\pi\omega np/c)\Lambda^{(1)}, \tag{A33}$$

where $\Lambda = \Lambda^{(1)}e^{ikz}$. Noting that $F = i\epsilon/2e^{ikz}$, the Bloch equations now are

$$\Lambda_t^{(1)} = (p\epsilon/h)\eta, \tag{A34}$$

$$\eta_t = (p\epsilon/h)\Lambda^{(1)}. \tag{A35}$$

We can eliminate the Bloch equations by the choice of an auxiliary variable $u(z,t)$:

$$-\eta = \cos u, \quad -\Lambda^{(1)} = \sin u, \quad p\epsilon/h = u_t. \tag{A36}$$

Then (A33) becomes

$$\left(\frac{\partial}{\partial z} + \frac{1}{c}\frac{\partial}{\partial t}\right)\frac{\partial u}{\partial t} = -\alpha_0 \sin u, \tag{A37}$$

the sine–Gordon equation, with

$$\alpha_0 = 4\pi^2 np^2/h\lambda. \tag{A38}$$

The propagation of a pulse in a nonlinear resonant medium is usually posed as a Goursat problem as follows: Given $\epsilon(0,t)$, $t > 0$, and given that at the initial time $t = 0$, the medium is in the unexcited ground state b, that is, $\Lambda(z,0) = \eta(z,0) + 1 = 0$, $z > 0$, find $\epsilon(z,t)$, $\Lambda^{(1)}(z,t)$, $\eta(z,t)$. (See Fig. 31.) Maxwell's equation (A33) tells us how ϵ changes along the characteristic $t - z/c = \text{const}$, whereas the Bloch equations tell us how $\Lambda^{(1)}$ and η are updated for increasing t at fixed z.

For the class of initial conditions

$$\frac{p}{h}\int_{-\infty}^{\infty} |\epsilon| dt < \infty,$$

the general solution has the following features: it consists of a finite number of kinks or 2π pulses,

$$u(z,t) = 4\tan^{-1}\exp\sqrt{\alpha_0}\eta(t - (z/c)(1 + c/\eta^2)),$$

$$(p\epsilon/h)(z,t) = 2\sqrt{\alpha_0}\eta\,\text{sech}\sqrt{\alpha_0}\eta(t - (z/c)(1 + c/\eta^2)), \tag{A39}$$

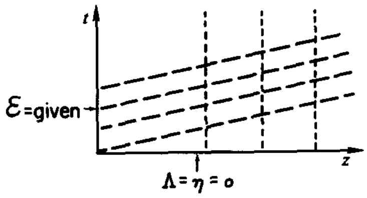

FIG. 31. Diagram for a pulse propagation as Goursat problem.

a finite number of 0π pulses or breathers,

$$u(z,t) = 4\tan^{-1}\left(\frac{n}{\xi}\,\text{sech}\,2\eta\left(t - \frac{z}{c}\left(1 + \frac{c}{4(\xi^2 + \eta^2)}\right)\right)\right)$$
$$\times \sin 2\xi\left(t - \frac{z}{c}\left(1 - \frac{c}{4(\xi^2 + \eta^2)}\right)\right), \tag{A40}$$

and radiation modes which have a continuous spectrum and are the nonlinear analog of solutions of the linearized equation (A37). Equation (A37) is in fact a soliton equation and the initial value problem posed above is separable and can in principle be solved exactly. The solitons (2π or 0π pulses) are so named because each carries an area

$$\frac{p}{h}\int_{-\infty}^{\infty} \epsilon\, dt = 2\pi \text{ or } 0$$

for all values of z.

3. The on resonance, inhomogeneous broadening case

If $\Gamma \neq 0$, then again the initial value problem can be solved exactly and again the initial profile decomposes into 0π and 2π pulses [very analogous to (A39) and (A40)] and into radiation.

The principal differences are as follows.

(i) The pulse is reshaped in a dimensionless distance of order $c\Gamma$ so that its area

$$A = \frac{p}{h}\int_{-\infty}^{\infty} \epsilon(z,t)dt$$

is an integer multiple of 2π (McCall–Hahn area theorem). Note that as the linewidth $\Gamma \to 0$, the area condition becomes a condition which the initial data must satisfy as was the case in the sine–Gordon equation discussed above.

(ii) The radiation is trapped in a dimensionless distance of order $(c\Gamma)^{-1}$ (Beer's law). Thus the medium is only "transparent" to the 0π or 2π pulses.

4. Homogeneous broadening

We now take account of losses in level population due to spontaneous emission. One can introduce these losses phenomenologically (as has been done in the literature) by postulating that the rate losses from levels a and b are linear and proportional to $\frac{1}{2}\gamma_a$ and $\frac{1}{2}\gamma_b$, respectively. This would mean that the losses of aa^*, bb^*, and $ab^* + a^*b$ would be γ_a, γ_b, and $\gamma_{ab} = \frac{1}{2}(\gamma_a + \gamma_b)$, respectively. However, things are not quite that simple [due to phonon interruptions by (defects) in solids and by atomic collisions in gases] and in fact it turns out that $\gamma_{ab} > \frac{1}{2}(\gamma_a + \gamma_b)$. In lasers, then, it is usual to treat γ_a, γ_b, and γ_{ab} as independent. In the situation discussed in this paper, the b state is the ground state and so we do not need the system to be pumped in order to keep the total population in states a and b constant. In fact, in order to keep the probability of finding the atom as either of the states a or b, we must have

$$aa^* + bb^* = 1, \tag{A41}$$

and thus aa^* and bb^* must satisfy

$$(aa^*)_t = -\gamma_a(aa^*), \quad (bb^*)_t = \gamma_a aa^*. \tag{A42}$$

This being the case, we find the net damping of $\eta = 2aa^* - 1$ to be

$$\eta_t = -\gamma_a(\eta + 1). \tag{A43}$$

It is not within the scope of this article to enter into the details of how the loss rate of the polarization is calculated. We will stipulate the loss can be described by adding $-\gamma_{ab}\Lambda$ to (A19).

With the addition of these losses, the Bloch equations are

$$\Lambda_t = -2i\zeta\Lambda - (2i/h)Fp\eta - \gamma_{ab}\Lambda, \tag{A44}$$

$$\eta_t = (ip/h)(F\Lambda^* - F^*\Lambda) - \gamma_a(\eta + 1), \tag{A45}$$

and the Maxwell equation is

$$\nabla^2 E - (1/c^2)E_{tt} = -(4\pi\omega^2/c^2)P \tag{A46}$$

where

$$P = \langle P_A \rangle \tag{A47}$$

and

$$P_A = \tfrac{1}{2}p(\Lambda e^{-i\omega t} + \Lambda^* e^{+i\omega t}). \tag{A48}$$

The sum $\langle\ \rangle$ is taken over the Lorentzian distribution

$$g(2\zeta) = \Gamma/\pi(\Gamma^2 + (2\zeta - 2\zeta_0)^2). \tag{A49}$$

Ohmic losses can be introduced by adding the damping term $-(4\pi/c)\sigma E_t$ to the left-hand side of (A46).

We are going to solve these equations in the limit where

$$|2\zeta_0| \gg \Gamma. \tag{A50}$$

The detuning is large with respect to the linewidth of inhomogeneous broadening and so we can take $\Gamma = 0$ and write (A47) as

$$P = \tfrac{1}{2}np(\Lambda e^{-i\omega t} + \Lambda^* e^{i\omega t}). \tag{A51}$$

The ζ in the Bloch equations in now ζ_0. The reader should realize that the further the system is detuned (from the linear viewpoint), the larger will be the necessary nonlinearity in order to have the response curve distort as so to produce effective nonlinear resonance.

We are also going to assume that the homogeneous linewidth γ is large with respect to the change of the electric field amplitude $F(z,t)$ at a fixed value of z. This means that we can treat $F(z,t)$ as a slowly varying function of t (slow with respect to times $1/\gamma$) in the Bloch equations (A44) and (A45) and then the population inversion η and polarization Λ can be computed by simply neglecting Λ_t and η_t. In this approximation, the polarization and population inversion follow the applied E field adiabatically.[12] We will continually return to query the uniform validity and self-consistency of this assumption.

It is now easy to show that

$$\eta = -\left(1 + \frac{4p^2}{h^2}\frac{\gamma_{ab}}{\gamma_a}\frac{FF^*}{\gamma_{ab}^2 + \Delta^2}\right)^{-1}, \tag{A52}$$

$$\Lambda = \frac{2ip}{hF}\left[(\gamma_{ab} - i\Delta)\left(1 + \frac{4p^2}{h^2}\frac{\gamma_{ab}}{\gamma_a}\frac{FF^*}{\gamma_{ab}^2 + \Delta^2}\right)\right]^{-1}, \tag{A53}$$

and

$$\nabla^2 E - \frac{1}{c^2}\frac{\partial^2 E}{\partial t^2} = \frac{2k\alpha_0}{\Delta}\biggl(Fe^{-i\omega t} + F^* e^{i\omega t}$$

$$- i\frac{\gamma_{ab}}{\Delta}Fe^{-i\omega t} + i\frac{\gamma_{ab}}{\Delta}F^* e^{+i\omega t}\biggr)$$

$$\times\biggl[1 + \frac{\gamma_{ab}^2}{\Delta^2}\biggl(1$$

$$+ \frac{4p^2}{h^2}\frac{\gamma_{ab}}{\gamma_a}\frac{FF^*}{\gamma^2 + \Delta^2}\biggr)\biggr]^{-1}, \tag{A54}$$

where we have set the detuning

$$2\zeta_0 + \omega_{ab} - \omega = -\Delta. \tag{A55}$$

In all cases, we will take

$$\Delta = O(10^3\gamma)\sec^{-1}, \quad \gamma_a \sim \gamma_{ab} \sim O(10^8)\sec^{-1}, \tag{A56}$$

and thus, for short lengths L_1 in the nonlinear medium $\alpha_0 L_1/\Delta - O(\pi)$, the attenuation is negligible. Let us rescale the electric field as

$$\frac{2p}{h}\frac{\gamma_{ab}}{\gamma_a(\gamma_{ab}^2 + \Delta^2)}\binom{E}{F} \to \sqrt{2}\binom{E}{F}, \tag{A57}$$

whence (A54) is

$$\nabla^2 E - \frac{1}{c^2}\frac{\partial^2 E}{\partial t^2} = \frac{2k\alpha_0}{\Delta}\frac{E}{1 + 2FF^*}, \tag{A58}$$

where

$$E(\mathbf{r},t) = F(\mathbf{r},t)e^{-i\omega t} + F^* e^{i\omega t}. \tag{A59}$$

The reason we took the carrier frequency ω greater than the two-level frequency is to ensure that we have a self-focusing rather than defocusing medium.

5. The ring cavity problem

This problem is posed as follows. Consider the situation shown in Fig. 1, in which a continuous input signal

$$E_{\text{in}} = A(\mathbf{x})e^{i(kz - \omega t)} + (*) \tag{A60}$$

[the electric field is scaled as in (A57)] enters a nonlinear medium through a partially transmitting mirror ($T \simeq 10\%$, $R \simeq 90\%$) at I. In the nonlinear medium, the electric field is written

$$E = G(\mathbf{x},t)e^{i(kz - \omega t)} + (*), \tag{A61}$$

where G changes slowly with respect to the transverse directions $\mathbf{x} = (x,y)$ and t. Note that $F = Ge^{ikz}$. Using the slowly varying envelope approximation in (A58) (the only term we neglect is $\partial^2 G/\partial t^2 \ll \omega\, \partial G/\partial t$), we have, after rescaling z as $\zeta = (2\alpha_0/\Delta)z$ and writing $\tau = t - z/c$,

$$2iG_\zeta + (\Delta/2\alpha_0 k)\nabla^2 G - G/(1 + 2GG^*) = 0. \tag{A62}$$

Equation (A62) tells us how an input signal $G(\zeta = 0, \tau = t, x)$ deforms along the characteristic $t - z/c = \tau = \text{const}$ due to the combined effects of diffraction $[(\Delta/2\alpha_0 k)\nabla^2 G]$ and nonlinearity. We call the parameter

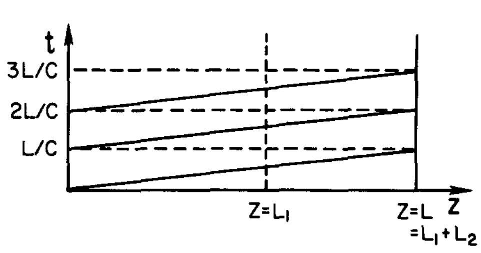

FIG. 32. Electric field propagation in the ring cavity in the *z-t* coordinates.

$$f = 2\alpha_0 k/\Delta w_0^2, \qquad \text{(A63)}$$

where w_0 is the transverse width of the entering pulse, the Fresnel number. The τ dependence is governed by the conditions applied to G at $z = 0$, the initial point of the nonlinear medium. We emphasize that the relative change of G along $t - z/c = \text{const}$ is $O(\Delta/\alpha_0)$, which may be large in comparison to $1/\gamma$, and does not invalidate the adiabatic assumption. As long as the change in G between characteristics is slow in comparison to time $1/\gamma$, the adiabatic approximation is fine.

After the signal has reached the end of the nonlinear medium at $z = L_1$ or $\zeta = 2\alpha_0 L_1/\Delta$, it is redirected back to the start by a pair of 100% reflecting mirrors at K and J and a pair of 90% reflecting mirrors at H and I. Here we make a rather artificial assumption; namely the pulse in the linear medium $IJKH$ undergoes no diffraction. In order to achieve this situation experimentally a very special lens device would have to be used. We shall ignore this difficulty and simply stipulate that the electric field after returning to H is equal to R times (two reflections) the electric field at I at the retarded time

$$E(z = L_1 + L_2, t, \mathbf{x}) = RE(z = L_1, t - L_2/c, \mathbf{x}). \qquad \text{(A64)}$$

The problem then is this. Given E_{in} beginning at $t = 0$, determine the electric field E at the beginning $z = 0$ of the nonlinear medium as $t \to \infty$. The special nature of the ring cavity allows us to replace the continuous time variable t by a discrete variable n. Why is this? Consider Fig. 32. The initial envelope at $z = 0$ for $0 < t < L/c$ (L/c is the signal round trip time) is given by

$$G_1(z = 0, \tau = t, \mathbf{x}) = \sqrt{T}A(\mathbf{x}) \qquad \text{(A65)}$$

and is independent of τ. Hence G_1 is independent of τ in the band $0 < t < L/C$. For $L/C < t < 2L/C$, there is a new initial condition

$$G_2(z = 0, \tau = t, \mathbf{x}) = \sqrt{T}A(x) + Re^{ikL}G_1(z = L_1, \mathbf{x}), \qquad \text{(A66)}$$

where the factor e^{ikL} arises from the phase shifts $e^{ikL_1 - i\omega(t - L_2/c)} = e^{ik(L_1 + L_2) - i\omega t}$ in the carrier wave; i.e.,

$$E_2(z = 0, \tau = t, \mathbf{x}) = \sqrt{T}A(x)e^{-i\omega t} + RE_1(z = L_1, \tau = t - L_2/c, \mathbf{x}).$$

But (A66) shows that $G_2(z = 0, \tau, \mathbf{x})$ is independent of τ. Therefore in each interval $nL/c < t < (n + 1)L/c$, G_{n+1} is independent of τ and

$$G_{n+1}(\mathbf{x}, z = 0) = \sqrt{T}A(x) + Re^{ikL}G_n(\mathbf{x}, z = L_1). \qquad \text{(A67)}$$

Our goal now is to find the function

$$\lim_{n \to \infty} G_n(\mathbf{x}, 0)$$

if the limit exists.

Equation (A67) is an infinite-dimensional map from a space of functions into itself. Given $G_n(\mathbf{x}, 0)$, one uses the partial differential equation

$$2iG_{n\zeta} + (1/f)\nabla^2 G_n - G_n/(1 + 2G_n G_n^*) = 0 \qquad \text{(A68)}$$

to determine $G_n(\mathbf{x}, 2\alpha_0 L_1/\Delta)$ from $G_n(\mathbf{x}, 0)$. The map may or may not have fixed points depending on the values of certain parameters like $a_0 = \text{Max}_x \sqrt{T}A(x)$ and $2\alpha_0 L_1/\Delta$. In fact, it exhibits all kinds of wild and wonderful behavior which we will discuss in Sec. VII.

We insert one small but important remark about the validity of the adiabatic approximation. It is clear that the electric field envelope $G(t,0)$ undergoes a discontinuity at the points $t = nL/c$, $n = 0,1,2,\ldots$, a discontinuity which is carried across the characteristics $t - z/c = nL/c$. Clearly in the neighborhood of these special characteristics, the field F in (2.59) is no longer slowly varying and the adiabatic approximation is invalid. However, one can show that E makes the transition across the discontinuity in a boundary layer of order $1/\gamma$ (γ stands for either γ_a or γ_{ab}). Therefore, as long as the round trip time L/c is much greater than the homogeneous broadening time γ^{-1}, the adiabatic approximation holds almost everywhere. It is important, however, to show that the solution which incorporates the detailed behavior of Δ, η across $nL/c - 1/\gamma < t < nL/c + 1/\gamma$ tends to the solution given in the previous pages in the limit $\gamma c/L \to 0$. The proof of this was given by Aceves *et al.*[1]

APPENDIX B: MATHEMATICAL DETAILS FOR THE LINEARIZED THEORY

In this appendix we describe the spectral theory of linearization (5.7c). This theory forms the foundation of the projections used to derive the reduced map on solitary wave parameters. First we change to real notation:

$$\widetilde{G}(y,z) = g(\theta;z)e^{i[(\lambda^2 - 1)/2]z}, \quad \theta = \lambda y,$$
$$g = U + iV, \quad \mathbf{g} = \begin{pmatrix} U \\ V \end{pmatrix}. \qquad \text{(B1a)}$$

Then linear equations (5.7c) take the real form

$$2\mathbf{g}_z = L\mathbf{g}, \quad L = \begin{pmatrix} 0 & L_- \\ -L_+ & 0 \end{pmatrix}, \qquad \text{(B1b)}$$

where the Schrödinger operators $L_\pm$ are defined by

$$L_- = -\partial_{\theta\theta} + 1 - (1/\lambda^2)[1 + N(S^2)],$$
$$L_+ = -\partial_{\theta\theta} + 1 - (1/\lambda^2)[1 + N(S^2) + 2N'(S^2)S^2]. \qquad \text{(B1c)}$$

The spectral theory of these Schrödinger operators is standard. They are of the form

$$L_\pm = -\partial_{\theta\theta} + V_\pm(\theta). \qquad \text{(B2)}$$

(See Figs. 33 and 34.) Letting $\theta \to \infty$ we see that the continuous spectrum of the operator L is determined by that of the skew adjoint constant coefficient operator

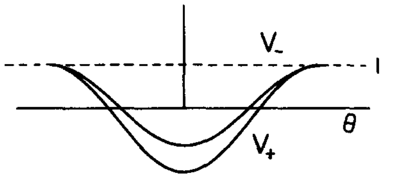

FIG. 33. The graphs of potentials $V_\pm(\theta)$ of the Schrödinger operators $L_\pm$. As $\theta\to\infty$, $V_\pm\to 1$.

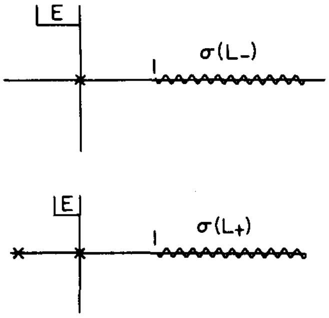

FIG. 34. The spectra of the operators $L_\pm$.

$$L_\infty = \begin{pmatrix} 0 & -\partial_{\theta\theta}+1 \\ \partial_{\theta\theta}-1 & 0 \end{pmatrix},$$

and is the union of two intervals on the imaginary axis, $(-i\infty, -i]\cup[i,i\infty)$. As for the discrete spectrum, we first observe that 0 is an eigenvalue of L with multiplicity 2; indeed, the null space of L is given by

$$\mathrm{N}(L) = \mathrm{span}\left\{\begin{pmatrix} S' \\ 0\end{pmatrix}, \begin{pmatrix} 0 \\ S\end{pmatrix}\right\}. \tag{B3a}$$

To see this, we calculate, using known properties of $L_\pm$,

$$\begin{aligned}\mathbf{n}\in\mathrm{N}(L) &\Leftrightarrow L\mathbf{n} = \mathbf{0}\\ &\Leftrightarrow L_-n_2 = 0,\quad L_+n_1 = 0\\ &\Leftrightarrow n_2 = 0,S \text{ and } n_1 = 0,S'\\ &\Leftrightarrow \mathbf{n} = c_1\begin{pmatrix} 0 \\ S\end{pmatrix} + c_2\begin{pmatrix} S' \\ 0\end{pmatrix}.\end{aligned}$$

Thus two z-independent solutions of (B1b) are

$$\mathbf{g}^{(1)} = \mathbf{n}^{(1)} = \begin{pmatrix} S' \\ 0\end{pmatrix},\quad \mathbf{g}^{(2)} = \mathbf{n}^{(2)} = \begin{pmatrix} 0 \\ S\end{pmatrix}.$$

These, of course, are equivalent to $\widetilde{G}^{(1)}, \widetilde{G}^{(2)}$ in (5.8) as generated by the symmetries of space and phase translation. In this real language the other two solutions $\widetilde{G}^{(3)}$ and $\widetilde{G}^{(4)}$, which grow linearly with z, are generated as follows. One considers the null space of L^2 which, since L is not skew adjoint, is not necessarily equal to the null space of L. In fact,

$$\begin{aligned}\mathbf{n}\in\mathrm{N}(L^2) &\Leftrightarrow L^2\mathbf{n} = \mathbf{0}\\ &\Leftrightarrow L\mathbf{n} = c_1\mathbf{g}^{(1)} + c_2\mathbf{g}^{(2)}\in\mathrm{N}(L)\\ &\Leftrightarrow \begin{pmatrix} 0 & L_- \\ -L_+ & 0\end{pmatrix}\begin{pmatrix} n_1 \\ n_2\end{pmatrix} = c_1\begin{pmatrix} 0 \\ S\end{pmatrix} + c_2\begin{pmatrix} S' \\ 0\end{pmatrix}\\ &\Leftrightarrow L_-n_2 = c_2S' \text{ and } L_+n_1 = -c_1S.\end{aligned}$$

First, we consider

$$L_-n_2 = S'.$$

Since $\mathrm{N}(L_-) = \mathrm{span}\{S\}$, this equation is solvable in $L^2(\mathbf{R})$ by the Fredholm alternative:

$$(S,L_-n_2) = (S,S'),\quad (L_-S,n_2) = (S,S'),$$
$$0 = \tfrac{1}{2}S^2|_{\theta=-\infty} = 0.$$

Indeed,

$$n_2 = -\tfrac{1}{2}\theta S,$$

as can be checked by differentiation. Similarly, since $\mathrm{N}(L_+) = \mathrm{span}\{S'\}$,

$$L_+n_1 = S$$

is solvable, and

$$n_1 = -\tfrac{1}{2}(\theta S_\theta + \lambda S_\lambda).$$

Thus we obtain

$$\begin{aligned}\mathrm{N}(L^2) &= \mathrm{N}(L)\cup\mathrm{span}\left\{\begin{pmatrix} 0 \\ \theta S\end{pmatrix}, \begin{pmatrix} \theta S_\theta + \lambda S_\lambda \\ 0\end{pmatrix}\right\}\\ &= \mathrm{span}\{\mathbf{n}^{(1)},\mathbf{n}^{(2)},\mathbf{n}^{(3)},\mathbf{n}^{(4)}\}.\end{aligned} \tag{B3b}$$

From these we generate two more solutions of the linear problem (B1b)

$$\mathbf{g}^{(3)} = z\mathbf{n}^{(1)} - \mathbf{n}^{(3)},\quad \mathbf{g}^{(4)} = z\mathbf{n}^{(2)} + \mathbf{n}^{(4)},$$

as can be quickly checked:

$$\begin{aligned}(2\partial_z - L)\mathbf{g}^{(3)} &= (2\partial_z - L)[z\mathbf{n}^{(1)} - \mathbf{n}^{(3)}]\\ &= 2\mathbf{n}^{(1)} - zL\mathbf{n}^{(1)} + L\mathbf{n}^{(3)}\\ &= 2\begin{pmatrix} S' \\ 0\end{pmatrix} + \begin{pmatrix} 0 & L_- \\ -L_+ & 0\end{pmatrix}\begin{pmatrix} 0 \\ \theta S\end{pmatrix}\\ &= \begin{pmatrix} L_-(\theta S) + 2S' \\ 0\end{pmatrix} = \begin{pmatrix} 0 \\ 0\end{pmatrix},\end{aligned}$$

with a similar calculation for $\mathbf{g}^{(4)}$.

Notice how elements in $\mathrm{N}(L^2)$ which are not members of $\mathrm{N}(L)$ generate solutions of the linear problem which grow linearly in "time" z. If $\mathrm{N}(L^3)$ contained functions which were not in the $\mathrm{N}(L^2)$, they would generate additional solutions of the linear problem which would grow quadratically in z. However, we have

$$\mathrm{N}(L^3) = \mathrm{N}(L^2), \tag{B3c}$$

because

$$\begin{aligned}\mathbf{n}\in\mathrm{N}(L^3) &\Leftrightarrow L^3\mathbf{n} = \mathbf{0}\\ &\Leftrightarrow L\mathbf{n} = C_1\mathbf{n}^{(1)} + C_2\mathbf{n}^{(2)} + C_3\mathbf{n}^{(3)} + C_4\mathbf{n}^{(4)}\\ &\Leftrightarrow L_-n_2 = C_1S' + C_4(\theta S_\theta + \lambda S_\lambda) \text{ and}\\ &\qquad -L_+n_1 = C_2S + C_3(\theta S).\end{aligned}$$

Now the second equation in this pair,

$$-L_+n_1 = C_2S + C_3(\theta S),$$

is solvable if and only if

$$(S',C_2S + C_3(\theta S)) = C_3(S',\theta S) = -C_3(S,S) = 0,$$

that is, if and only if $C_3 = 0$. The first equation

$$L_-n_2 = C_1S' + C_4(\theta S_\theta + \lambda S_\lambda)$$

is solvable

$$\Leftrightarrow (S, C_1 S' + C_4(\theta S_\theta + \lambda S_\lambda)) = 0$$
$$\Leftrightarrow C_4[-\tfrac{1}{2}(S,S) + (S,\lambda S_\lambda)] = 0.$$

We would like to conclude that $C_4 = 0$. For the Kerr case, this follows from the fact $S(\theta;\lambda) = \lambda \operatorname{sech} \theta$:

$$\begin{aligned} C_4\Big[-\frac{1}{2}(S,S) + (S,\lambda S_\lambda)\Big] &= C_4\Big[\lambda^2\Big(1-\frac{1}{2}\Big)\int_{-\infty}^{\infty} \operatorname{sech}^2\theta\, d\theta\Big] \\ &= 0 \Leftrightarrow C_4 = 0. \end{aligned}$$

For the case of saturable nonlinearity, we have checked numerically that $[-\frac{1}{2}(S,S) + (S,\lambda S_\lambda)] \neq 0$, and hence $C_4 = 0$. This shows $\mathrm{N}(L^3) = \mathrm{N}(L^2)$ for the two cases studied here.

Summarizing the situation, we are constructing a complete set of solutions of linear problem (B1b),

$$2\mathbf{g}_z = L\mathbf{g},$$

by studying the spectral theory of the operator L. So far, we understand that the continuous spectrum of L resides on the imaginary axis and therefore yields solutions of (B1b) that are bounded in x and oscillate in z. In addition 0 is in the point spectrum of L and generates exactly four solutions of (B1b), two of which are independent of z and the other two grow linearly with z. Actually, since L is not skew adjoint, we must develop a biorthogonal expansion based upon eigenfunctions of $L^\dagger$ as well as of L. The representation

$$L^\dagger = \begin{pmatrix} 0 & -L_+ \\ L_- & 0 \end{pmatrix}$$
$$L^\dagger = JLJ, \quad J = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix},$$

shows that if ψ is an eigenfunction of L with eigenvalue E, then $J\psi$ is an eigenfunction of $L^\dagger$ with eigenvalue $-E$. In particular, we have

$$\mathrm{N}(L^{\dagger 2}) = J\mathrm{N}(L^2). \tag{B3d}$$

We can label the elements of these null spaces as follows:

$$\begin{aligned} &\mathrm{N}(L^{\dagger 2}) = \operatorname{span}\{\mathbf{a}^{(i)},\ i = 1,2,3,4\}, \\ &\mathrm{N}(L^2) = \operatorname{span}\{\mathbf{A}^{(i)},\ i = 1,2,3,4\}, \end{aligned} \tag{B4a}$$

where

$$\begin{aligned} &\mathbf{a}^{(1)} = \begin{pmatrix} S \\ 0 \end{pmatrix}, \quad \mathbf{A}^{(1)} = \gamma_A \begin{pmatrix} (1/\lambda)\theta S_\theta + S_\lambda \\ 0 \end{pmatrix}, \\ &\mathbf{a}^{(2)} = \begin{pmatrix} 0 \\ -S_\theta \end{pmatrix}, \quad \mathbf{A}^{(2)} = \gamma_B \begin{pmatrix} 0 \\ \theta S \end{pmatrix}, \\ &\mathbf{a}^{(3)} = \begin{pmatrix} \theta S \\ 0 \end{pmatrix}, \quad \mathbf{A}^{(3)} = \gamma_B \begin{pmatrix} -S_\theta \\ 0 \end{pmatrix}, \\ &\mathbf{a}^{(4)} = \begin{pmatrix} 0 \\ (1/\lambda)\theta S_\theta + S_\lambda \end{pmatrix}, \quad \mathbf{A}^{(4)} = \gamma_A \begin{pmatrix} 0 \\ S \end{pmatrix}. \end{aligned} \tag{B4b}$$

If we choose the constants γ_A, γ_B as

$$\begin{aligned} &\gamma_A = \frac{2}{(-1/\lambda + \partial/\partial\lambda)(S,S)}, \\ &\gamma_B = 2/(S,S), \end{aligned} \tag{B4c}$$

these are paired in a biorthogonal fashion

$$(\mathbf{a}^{(i)}, \mathbf{A}^{(j)}) = \delta_{ij}. \tag{B4d}$$

Again consider the linear problem (B1b),

$$2\mathbf{g}_z = L\mathbf{g},$$

and define the quadratic form as

$$Q(\mathbf{g}) = \langle \mathbf{g}, JL\mathbf{g} \rangle = (L_+ g_1, g_1) + (L_- g_2, g_2). \tag{B5}$$

If $\mathbf{g}(z)$ satisfies the linear problem (B1b), then the quadratic form $Q(\mathbf{g}(z))$ is an invariant:

$$\begin{aligned} \frac{d}{dz} Q[\mathbf{g}(z)] &= 2(L_+ g_1, \dot g_1) + 2(L_- g_2, \dot g_2) \\ &= (L_+ g_1, L_- g_2) + (L_- g_2, -L_+ g_1) = 0. \end{aligned}$$

This invariant quadratic form $Q[\mathbf{g}]$ is similar to the H_1 norm of $\mathbf{g}$, as integration by parts shows:

$$\begin{aligned} Q(\mathbf{g}) &= (L_+ g_1, g_1) + (L_- g_2, g_2) \\ &= ([-\partial_{\theta\theta} + 1 - (1/\lambda^2)U_+] g_1, g_1) \\ &\quad + ([-\partial_{\theta\theta} + 1 - (1/\lambda^2)U_-] g_2, g_2) \\ &= (g_1', g_1') + (g_1, g_1) - (1/\lambda^2)(U_+ g_1, g_1) \\ &\quad + (g_2', g_2') + (g_2, g_2) - (1/\lambda^2)(U_- g_2, g_2), \end{aligned} \tag{B6}$$

where the postive functions $U_\pm$ are defined by

$$\begin{aligned} &U_- = 1 + N(S^2), \\ &U_+ = 1 + N(S^2) + 2N'(S^2)S^2. \end{aligned} \tag{B7}$$

We would like to use this quadratic invariant $Q(\mathbf{g})$ as a norm; unfortunately, it is not positive definite due to the attractive potentials $-U_\pm$. Indeed, let $\mathbf{g}_b = (b,0)$, where b is the negative energy ground state of L_+ with eigenvalue $-|E_0|$. Then

$$Q(\mathbf{g}_b) = (L_+ b, b) = -|E_0|(b,b) < 0.$$

In a related consideration, we compute this quadratic form on the four solutions $\mathbf{g}^{(j)}$ of the linear problem (B1b):

$$\begin{aligned} Q(\mathbf{g}^{(1)}) &= (L_+ S', S') = (0, S') = 0, \\ Q(\mathbf{g}^{(2)}) &= (L_- S, S) = (0, S) = 0, \\ Q(\mathbf{g}^{(3)}(z)) &= Q(\mathbf{g}^{(3)}(z=0)) = Q(-\mathbf{n}^{(3)}) \\ &= (L_- \theta S, \theta S) = (-2S', \theta S) = (S,S), \\ Q(\mathbf{g}^{(4)}(z)) &= Q(\mathbf{g}^{(4)}(0)) = Q(\mathbf{n}^{(4)}) \\ &= (L_+(\theta S_\theta + \lambda S_\lambda), \theta S_\theta + \lambda S_\lambda) \\ &= -2(S, \theta S_\theta + \lambda S_\lambda) \\ &= (S,S) - 2(S, \lambda S_\lambda) < 0. \end{aligned}$$

Notice in particular that $\mathbf{g}^{(1)}$, $\mathbf{g}^{(2)}$, and $\mathbf{g}^{(4)}$ are badly behaved with respect to $Q(\cdot)$.

In order to use the quadratic form $Q(\cdot)$ as a norm, we must work in a slightly smaller space than H_1—one orthogonal to $\mathrm{N}(L^{\dagger 2})$. We decompose H_1 as follows:

$$H_1 = \mathbf{M} \oplus \mathrm{N}(L^2), \quad \mathbf{M} = [H_1 \cap (\mathrm{N}(L^{\dagger 2}))^\perp]. \tag{B8}$$

That is, for each $\mathbf{g} \in H_1$, we write

$$\mathbf{g} = \sum_{i=1}^{4} \alpha_i \mathbf{A}^{(i)} + \mathbf{g}_M,$$

where

$$\alpha_i = \langle \mathbf{a}^{(i)},\mathbf{g}\rangle,$$

which implies

$$\mathbf{g}_M \perp \mathrm{N}(L^{\dagger 2}).$$

Because the splitting (B8) was defined using the biorthogonal pairing (B4a), L acts invariantly with respect to this splitting,

$$\begin{aligned} &L\colon\ \mathrm{N}(L^2)\to\mathrm{N}(L^2)\\ &L^\dagger\colon\ \mathrm{N}(L^{\dagger 2})\to\mathrm{N}(L^{\dagger 2}),\\ &L\colon\ \mathbb{M}\to\mathbb{M}, \end{aligned}\tag{B9}$$

as can be quickly checked:

$$\mathbf{g}\in\mathbb{M}\Rightarrow\langle\mathbf{a}^{(i)},\mathbf{g}\rangle = 0$$

$$\Rightarrow\langle\mathbf{a}^{(i)},L\mathbf{g}\rangle = \langle L^\dagger\mathbf{a}^{(i)},\mathbf{g}\rangle = \left\langle\sum_{j=1}^{4} c_{ij}\mathbf{a}^{(j)},\mathbf{g}\right\rangle = 0.$$

This invariant action of L as described by (B9) guarantees that L does not couple the subspaces $\mathbb{M}$ and $\mathrm{N}(L^2)$ and that splitting (B8) is consistent with the z evolution. That is, the initial value problem,

$$2\mathbf{g}_z = L\mathbf{g},\quad \mathbf{g}|_{z=0} = \mathbf{h}\in H_1,\tag{B10}$$

can be split into two completely decoupled problems,

$$2\mathbf{g}_z^{(\mathrm{M})} = L\mathbf{g}^{(\mathrm{M})},\quad \mathbf{g}^{(\mathrm{M})}|_{z=0} = \mathbf{h}^{(\mathrm{M})},\tag{B11a}$$

$$2\mathbf{g}_z^{(\mathrm{N})} = L\mathbf{g}^{(\mathrm{N})},\quad \mathbf{g}^{(\mathrm{N})}|_{z=0} = \mathbf{h}^{(\mathrm{N})},\tag{B11b}$$

where $\mathbf{g}^{(\mathrm{M})}$ and $\mathbf{h}^{(\mathrm{M})}\in\mathbb{M}$, $\mathbf{g}^{\mathrm{N}}$ and $\mathbf{h}^{(\mathrm{N})}\in\mathrm{N}(L^2)$,

$$\mathbf{g} = \mathbf{g}^{(\mathrm{M})} + \mathbf{g}^{(\mathrm{N})},\quad \mathbf{h} = \mathbf{h}^{(\mathrm{N})} + \mathbf{h}^{(\mathrm{N})},$$

$$\mathbf{h}^{(\mathrm{N})} = \sum_{i=1}^{4}\langle\mathbf{a}^{(i)},\mathbf{h}\rangle\mathbf{A}^{(i)},\quad \mathbf{h}^{(\mathrm{M})} = \mathbf{h} - \mathbf{h}^{(\mathrm{N})}.$$

Because the dimension of $\mathrm{N}(L^2)$ is 4, Eq. (B11b) is really a fourth-order ordinary differential equation. Here this finite-dimensional system is trivial:

$$\begin{aligned}
&\mathbf{g}^{(\mathrm{N})}(z) = \sum_{i=1}^{4}\alpha_i(z)\mathbf{A}^{(i)}\\
&\quad\Rightarrow\sum_{i=1}^{4}2\dot\alpha_i(z)\mathbf{A}^{(i)} = \sum_{i=1}^{4}\alpha_i(z)L\,\mathbf{A}^{(i)},\\
&\quad\Rightarrow 2\dot\alpha_1 = 0,\quad 2\dot\alpha_2 = 0,\quad 2\dot\alpha_3 = \sum_{i=1}^{4}\alpha_i(z)\langle L^{+}\mathbf{a}^{(3)},\mathbf{A}^{(i)}\rangle = \sum_{i=1}^{4}2\alpha_i(z)\langle\mathbf{a}^{(2)},\mathbf{A}^{(i)}\rangle = 2\alpha_2,\\
&\qquad 2\dot\alpha_4 = \sum_{i=1}^{4}\alpha_i(z)\langle L^{+}\mathbf{a}^{(4)},\mathbf{A}^{(i)}\rangle = \sum_{i=1}^{4}\frac{2}{\lambda}\alpha_i(z)\langle\mathbf{a}^{(1)},\mathbf{A}^{(i)}\rangle = \frac{2}{\lambda}\alpha_1,\\
&\quad\Rightarrow\dot\alpha_1 = 0,\quad \dot\alpha_2 = 0,\quad \dot\alpha_3 = \alpha_2,\quad \dot\alpha_4 = \alpha_1/\lambda,\quad \alpha_i(0) = \langle\mathbf{a}^{(i)},\mathbf{h}\rangle,\quad i = 1,2,3,4,\\
&\quad\Rightarrow\mathbf{g}^{(\mathrm{N})}(\zeta) = \langle\mathbf{a}^{(1)},\mathbf{h}\rangle\mathbf{A}^{(1)} + \langle\mathbf{a}^{(2)},\mathbf{h}\rangle\mathbf{A}^{(2)} + [\langle\mathbf{a}^{(2)},\mathbf{h}\rangle z + \langle\mathbf{a}^{(3)},\mathbf{h}\rangle]A^{(3)} + [\langle\mathbf{a}^{(1)},\mathbf{h}\rangle z/\lambda + \langle\mathbf{a}^{(4)},\mathbf{h}\rangle]A^{(4)}.
\end{aligned}\tag{B12}$$

On $\mathbb{M}$, Eq. (B11a) is still a partial differential equation; however, we can control $\mathbf{g}^{(\mathrm{M})}(z)$ using the quadratic invariant. For, on $\mathbb{M}$, we have the following.[10]

Theorem: $\exists C_1, C_2 > 0$, $\forall\mathbf{g}\in\mathbb{M}$,

$$0\leqslant C_1\|\mathbf{g}\|_1^2 < Q(\mathbf{g}) < C_2\|\mathbf{g}\|_1^2.\tag{B13}$$

This is an extremely useful result. For example, it can be used to establish the linearized stability of the solitary wave G_s to perturbations of the initial data. By definition of linearized stability, one must control $\mathbf{g}(z)$ $\forall z$, where $\mathbf{g}(z)$ satisfies the initial value problem

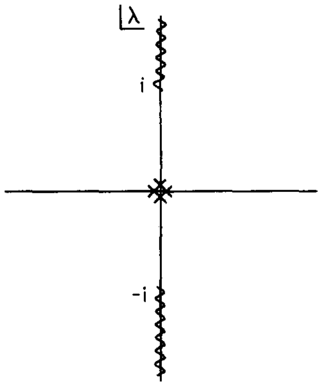

FIG. 35. The spectrum of L: continuous spectrum lies in $(-i\infty,-i]\cup[i,i\infty)$ and a point spectrum at $\lambda = 0$.

$$2\mathbf{g}_z = L\mathbf{g},\quad \mathbf{g}(z=0) = \mathbf{h}\in H_1.$$

Using splitting (B9),

$$\mathbf{g} = \mathbf{g}^{(\mathrm{N})} + \mathbf{G},\quad \mathbf{h} = \mathbf{h}^{(\mathrm{N})} + \mathbf{H},$$

where $\mathbf{G},\mathbf{H}\in\mathbb{M}$ and satisfy

$$2\mathbf{G}_z = L\,\mathbf{G},\quad \mathbf{G}_{(z=0)} = \mathbf{H}\in\mathbb{M}.$$

We use (B13) to control $\mathbf{G}(z)$ $\forall z$: Consider any $\epsilon > 0$, and assume

$$\|\mathbf{H}\|_1^2 < (C_1/C_2)\epsilon.$$

Then

$$C_1\epsilon > C_2\|\mathbf{H}\|_1^2 > Q(\mathbf{G}(0)) = Q(\mathbf{G}(z)) > C_1\|\mathbf{G}(z)\|_1^2$$

$$\Rightarrow\|\mathbf{G}(z)\|_1^2 < \epsilon.$$

Thus the only growth of $\mathbf{g}(z)$ comes from the four-dimensional system for $\mathbf{g}^{(\mathrm{N})}$ and is at most linear in z as (B12) shows.

Much more can be obtained from estimate (B13). It shows that on M, the quadratic invariant may be used to define a norm which is equivalent to the H_1 norm. Now $\mathbb{M}$ can be turned into a Hilbert space, with inner product

$$\langle\mathbf{g},\mathbf{h}\rangle_{\mathrm{M}} = \langle\mathbf{g},JL\,\mathbf{h}\rangle.\tag{B14}$$

With respect to this "energy" inner product, L is skew adjoint:

$$\langle \mathbf{g}, L\,\mathbf{h}\rangle_{\mathbf{M}} = -\langle L\mathbf{g}, \mathbf{h}\rangle_{\mathbf{M}}. \tag{B15}$$

Thus on **M** the spectrum of L is purely imaginary. A completeness theorem and an eigenfunction expansion are at our disposal. The spectrum of L is summarized in Fig. 35.

[1] A. Aceves, H. Adachihara, C. Jones, J. C. Lerman, D. W. McLaughlin, J. V. Moloney, and A. C. Newell, Physica D **18**, 85 (1986).

[2] (a) *Optical Bistability*, Vol. I, edited by C. M. Bowden, M. Ciftan, and H. Robl (Plenum, New York, 1981); Vol. II, edited by C. M. Bowden, H. M. Gibbs, and S. L. McCall (Plenum, New York, 1984); Vol. III, edited by H. M. Gibbs, P. Mandel, M. Peyghambarian, and S. D. Smith (Springer, Berlin, 1986); (b) H. M. Gibbs, *Optical Bistability—Controlling Light with Light* (Academic, New York, 1985).

[3] D. W. McLaughlin, J. V. Moloney, and A. C. Newell, Phys. Rev. Lett. **51**, 75 (1983).

[4] D. W. McLaughlin, J. V. Moloney, and A. C. Newell, Phys. Rev. Lett. **54**, 681 (1985).

[5] D. W. McLaughlin, J. V. Moloney, and A. C. Newell, in *Chaos in Nonlinear Dynamical Systems*, edited by J. Chandra (SIAM, Philadelphia, 1984).

[6] (a) J. V. Moloney, IEEE J. Quant. Elect. **21**, 1393 (1985); (b) C. Jones, D. W. McLaughlin, J. V. Moloney, and A. C. Newell, "Two dimensional effects in optical ring cavity," in preparation.

[7] J. V. Moloney, Phys. Rev. A **33**, 4061 (1986).

[8] (a) J. V. Moloney, Opt. Commun. **45**, 435 (1984); (b) S. M. Hammel, C. K. R. T. Jones, and J. V. Moloney, J. Opt. Soc. Am. B **2**, 552 (1985).

[9] (a) K. Ikeda, H. Daido, and O. Akimoto, Phys. Rev. Lett. **45**, 709 (1980); (b) H. J. Carmichael, R. R. Snapp, and W. C. Schieve, Phys. Rev. A **26**, 3408 (1982), and references therein; (c) P. Mandel and R. Kapral, Opt. Commun. **47**, 151 (1983).

[10] M. Weinstein, SIAM J. Math Anal. **16**, 472 (1985).

[11] R. Meinel and G. Neugebauer, Phys. Lett. A **100**, 467 (1984).

[12] H. Haken and H. Sauermann, Z. Phys. **9** (1963).

Exact solutions for wave equations of two-layered media with smooth transition

George Bluman
Department of Mathematics and Institute of Applied Mathematics, University of British Columbia, Vancouver, Canada V6T 1Y4
Sukeyuki Kumei
The Faculty of Textile Science and Technology, Shinshu University, 3-15-1 Tokida, Ueda, Nagano ken, Japan

(Received 5 January 1987; accepted for publication 13 August 1987)

The wave equation $c^2(x)u_{xx} - u_{tt} = 0$ is solved for wave speeds $c(x)$ corresponding to two-layered media with smooth transition from layer to layer. The wave speed $c(x)$ has four free parameters to fit a given medium. Solutions are constructed from invariant solutions of a related system of first-order partial differential equations that admit a four-parameter symmetry group. These solutions are superposed to solve general initial value problems for data with compact support; the computation of the superposition coefficients uses elementary Fourier analysis. Solutions are illustrated for various initial conditions.

I. INTRODUCTION

In a previous paper[1] we classified all wave equations of the form

$$c^2(x)u_{xx} - u_{tt} = 0, \tag{1.1}$$

which are solvable by group theoretical methods. In particular, we showed that the system of partial differential equations

$$v_t = u_x, \quad u_t = c^2(x)v_x \tag{1.2}$$

equivalent to Eq. (1.1), admits a maximal four-parameter Lie group of point transformations if and only if the wave speed $c(x)$ satisfies the ordinary differential equation

$$cc'(c/c')'' = \text{const} = \mu. \tag{1.3}$$

If $\mu = 0$, the solution of Eq. (1.3) reduces to either $c(x) = e^x$ or $c(x) = x^A$, where A is an arbitrary constant. In Ref. 1 we constructed the corresponding invariant solutions of (1.2).

If $\mu \neq 0$, Eq. (1.3) reduces to one of the following four standard forms[1]:

$\boxed{\mu = 1}$

$$c' = \nu^{-1}\sin(\nu \log c); \tag{1.4}$$

$$c' = \nu^{-1}\sinh(\nu \log c); \tag{1.5}$$

$$c' = \log c; \tag{1.6}$$

$\boxed{\mu = -1}$

$$c' = \nu^{-1}\cosh(\nu \log c); \tag{1.7}$$

where $\nu \neq 0$ is an arbitrary constant. Solutions of (1.1) and (1.2) are discussed in Ref. 2 for $c(x)$ satisfying (1.5) or (1.7) with $\nu = \frac{1}{2}$.

If $c(x) = \phi(x,\nu)$ is a solution of any one of the equations (1.4)–(1.7) then the corresponding general solution of Eq. (1.3) is given by

$$c(x) = K\phi(Lx + M,\nu), \tag{1.8}$$

where $K^2L^2 = |\mu|$ for any constants $\{L,M,\nu\}$.

For each of the equations (1.4)–(1.7) solutions $c(x)$ are monotone functions of x; $c(x)$ is bounded on $(-\infty,\infty)$ if and only if $c(x)$ satisfies Eq. (1.4). Such a bounded $c(x)$ has a smooth simple jump (cf. Fig. 1). This corresponds to wave propagation in a two-layered stratified medium with a smooth transition from layer to layer.

In the rest of this paper we construct various invariant solutions of system (1.2) and hence solutions of (1.1), where the wave speed $c(x)$ satisfies (1.4); without loss of generality $\nu > 0$. We show how to solve general initial value problems by a superposition of these invariant solutions. We illustrate our results by solving initial value problems for initial humps of varying shape and location.

II. PROPERTIES OF $c(x)$

Say $c(x)$ solves (1.4). Then $|c'(x)| \leqslant 1/\nu$, and $c'(x) = 0$ if and only if

$$c(x) = e^{k\pi/\nu}, \quad k = 0, \pm 1, \pm 2,\dots . \tag{2.1}$$

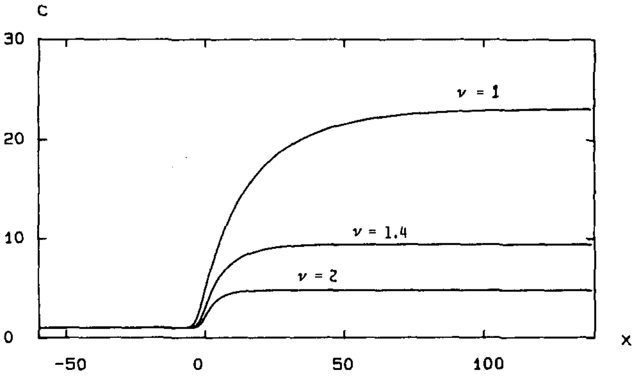

FIG. 1. Profile of $c(x) = \Phi(x,\nu)$.

Now consider the region where

$$1 < c(x) < e^{\pi/\nu}. \tag{2.2}$$

In this strip $c'(x) > 0$ and the inflection point $x = x^*$ occurs where $c(x^*) = e^{\pi/2\nu}$; $c'(x^*) = 1/\nu$. Equation (1.4) leads to

$$\lim_{\epsilon \to 0^+} \int_{e^{\pi/2\nu}}^{e^{\pi/\nu} - \epsilon} \frac{dc}{\sin(\nu \log c)} = +\infty, \tag{2.3}$$

and

$$\lim_{\epsilon \to 0^+} \int_{1+\epsilon}^{e^{\pi/2\nu}} \frac{dc}{\sin(\nu \log c)} = +\infty. \tag{2.4}$$

Hence it follows that $\lim_{x \to +\infty} c(x) = e^{\pi/\nu}$, $\lim_{x \to -\infty} c(x) = 1$. Thus $c = 1$, $c = e^{\pi/\nu}$ are horizontal asymptotes for $c(x)$ in the strip (2.2). Since Eq. (1.4) is invariant under translation in x, without loss of generality we can set $x^* = 0$.

Now let

$$c(x) = \Phi(x,\nu) \tag{2.5}$$

be the solution of Eq. (1.4) with properties

$$\lim_{x \to -\infty} \Phi(x,\nu) = 1, \tag{2.6}$$

$$\lim_{x \to +\infty} \Phi(x,\nu) = e^{\pi/\nu}, \tag{2.7}$$

$$\Phi(0,\nu) = e^{\pi/2\nu}. \tag{2.8}$$

One can show that Eq. (1.4) has solutions

$$c(x) = e^{-n\pi/\nu}\Phi((-1)^n e^{n\pi/\nu}x,\nu) \tag{2.9}$$

on the horizontal strip

$$e^{n\pi/\nu} < c(x) < e^{(n+1)\pi/\nu}, \tag{2.10}$$

$n = 0, \pm 1, \pm 2,\ldots$.

From property (1.8) it follows that each strip solution leads to the same general solution of Eq. (1.3). Thus from now on we will only consider the solution $c(x) = \Phi(x,\nu)$ of Eq. (1.4).

Graphs of $\Phi(x,\nu)$ and $(d/dx)\Phi(x,\nu)$ are given in Figs. 1 and 2, respectively, for $\nu = 1$, 1.4, 2.

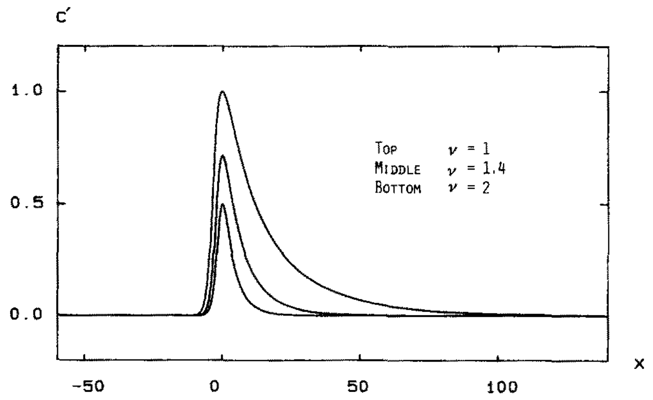

FIG. 2. $c' = \Phi'(x,\nu)$.

Here $\Phi(x,\nu)$ has asymptotic properties,

$$\Phi(x,\nu) = 1 + C^-(\nu)e^x + o(e^x) \quad \text{as } x \to -\infty, \tag{2.11}$$

$$\Phi(x,\nu) = e^{\pi/\nu}[1 - C^+(\nu)e^{-(e^{-\pi/\nu})x}] + o(e^{-(e^{-\pi/\nu})x}) \quad \text{as } x \to +\infty, \tag{2.12}$$

$$\Phi(x,\nu) = e^{\pi/2\nu} + x/\nu + o(x^2) \quad \text{as } x \to 0, \tag{2.13}$$

for some positive constants $\{C^-(\nu), C^+(\nu)\}$.

To obtain a bounded monotonically increasing solution $c(x)$ of Eq. (1.3) with the properties

$$\lim_{x \to -\infty} c(x) = c_1, \tag{2.14}$$

$$\lim_{x \to +\infty} c(x) = c_2, \tag{2.15}$$

and

$$\max_{x \in (-\infty,\infty)} c'(x) = m, \tag{2.16}$$

where $\{c_1, c_2, m\}$ are arbitrary positive constants with $0 < c_1 < c_2$, we set in (1.8),

$$K = c_1, \quad L = (m/c_1)\nu^*, \\ \nu = \nu^* = \pi(\log c_2/c_1)^{-1}. \tag{2.17}$$

The general solution of Eq. (1.3) satisfying (2.14)–(2.16) is

$$c(x) = c_1\Phi((m/c_1)\nu^* x + M,\nu^*), \tag{2.18}$$

where M is an arbitrary constant.

The width of the transition region in x is $O((c_2 - c_1)/m)$. Since $\Phi(x,\nu)$ exponentially approaches its horizontal asymptotes, a wave speed $c(x)$, represented by (2.18), effectively approximates a two-layered medium. The transition between layers can be as abrupt as one wishes.

III. INVARIANCE PROPERTIES OF SYSTEM (1.2)

As shown in Ref. 1, when $c(x)$ satisfies (1.3) for $\mu > 0$, the system (1.2) admits the four-parameter $\{p,q,r,s\}$ Lie group of point transformations

$$\begin{aligned} X &= x + \epsilon\xi(x,t) + O(\epsilon^2), \\ T &= t + \epsilon\tau(x,t) + O(\epsilon^2), \\ U &= u + \epsilon[i(x,t)u + j(x,t)v] + O(\epsilon^2), \\ V &= v + \epsilon[k(x,t)v + l(x,t)u] + O(\epsilon^2), \end{aligned} \tag{3.1}$$

where in terms of

$$\beta(t) = pe^t - qe^{-t}, \tag{3.2}$$

$\{\xi,\tau,i,j,k,l\}$ are given by

$$\begin{aligned} \xi &= 2\beta'(t)[c(x)/c'(x)], \\ \tau &= 2\beta(t)[(c(x)/c'(x))' - 1] + r, \\ i &= \beta'(t)[2 - (c(x)/c'(x))'] + s, \\ j &= -\beta(t)[c(x)/c'(x)], \\ k &= -\beta'(t)[c(x)/c'(x)]' + s, \\ l &= -\beta(t)[1/c(x)c'(x)]. \end{aligned} \tag{3.3}$$

The group generators for the parameters $\{p,q,r,s\}$, respectively, are

$$
\begin{aligned}
L_p &= e^t\left[\frac{2c}{c'}\frac{\partial}{\partial x}+2\left[\left(\frac{c}{c'}\right)'-1\right]\frac{\partial}{\partial t}\right.\\
&\quad+\left[\left[2-\left(\frac{c}{c'}\right)'\right]u-\frac{c}{c'}v\right]\frac{\partial}{\partial u}\\
&\quad\left.-\left[\left(\frac{c}{c'}\right)'v+\frac{1}{cc'}u\right]\frac{\partial}{\partial v}\right],\\
L_q &= e^{-t}\left[\frac{2c}{c'}\frac{\partial}{\partial x}+2\left[1-\left(\frac{c}{c'}\right)'\right]\frac{\partial}{\partial t}\right.\\
&\quad+\left[\left[2-\left(\frac{c}{c'}\right)'\right]u+\frac{c}{c'}v\right]\frac{\partial}{\partial u}\\
&\quad\left.-\left[\left(\frac{c}{c'}\right)'v-\frac{1}{cc'}u\right]\frac{\partial}{\partial v}\right],\\
L_r &= \frac{\partial}{\partial t},\quad L_s=u\frac{\partial}{\partial u}+v\frac{\partial}{\partial v}.
\end{aligned}
\tag{3.4}
$$

The commutators of the Lie algebra are

$$
\begin{aligned}
&[L_p,L_q]=-8\nu^2L_r,\quad [L_r,L_p]=L_p,\\
&[L_r,L_q]=-L_q,\quad [L_p,L_s]=[L_q,L_s]=[L_r,L_s]=0.
\end{aligned}
\tag{3.5}
$$

The global transformation generated by (3.1)–(3.3) is found by solving the characteristic equations

$$
\begin{aligned}
\frac{dX}{\xi(X,T)} &= \frac{dT}{\tau(X,T)}\\
&= \frac{dU}{i(X,T)U+j(X,T)V}\\
&= \frac{dV}{k(X,T)V+l(X,T)U}=d\epsilon,
\end{aligned}
\tag{3.6}
$$

where

$$X=x,\quad T=t,\quad U=u,\quad V=v,\quad \text{at }\epsilon=0. \tag{3.7}$$

The global transformation for $r\neq0$ is obtained from the global transformation for $r=0$ by letting $t\to t+r$. Without loss of generality we set $r=0$.

Now let

$$Y=\nu\log c(X),\quad \gamma=4pq=(\beta')^2-\beta^2. \tag{3.8}$$

Then the resulting implicit global transformation is

$$
\begin{aligned}
&z=\beta(T)\sin Y,\\
&[[c(X)]^{-1/2}U+[c(X)]^{1/2}V]^2\\
&\quad=[Ae^{2\epsilon s}\sin Y][\beta(T)\cos Y+\beta'(T)],\\
&[[c(X)]^{-1/2}U-[c(X)]^{1/2}V]^2\\
&\quad=[Be^{2\epsilon s}\sin Y][\beta(T)\cos Y-\beta'(T)];
\end{aligned}
\tag{3.9}
$$

and

$$
\begin{aligned}
&(1/\nu\sqrt{-\gamma})\log|\cos Y+(1/\sqrt{-\gamma})\beta'(T)\sin Y|\\
&\quad=E-2\epsilon\quad\text{for }\gamma<0,
\end{aligned}
\tag{3.10}
$$

$$(1/\nu\sqrt{\gamma})\arctan\left[\frac{\sqrt{\gamma}}{\beta'(T)}\cot Y\right]=E-2\epsilon\quad\text{for }\gamma>0. \tag{3.11}$$

The integration constants $\{z,A,B,E\}$ are expressed in terms of $\{x,t,u,v\}$ by using the initial condition (3.7). Without loss of generality $\gamma=1$ if $\gamma>0$, $\gamma=-1$ if $\gamma<0$.

Now we construct invariant solutions of system (1.2) for $r=0$. Let

$$y=\nu\log c(x). \tag{3.12}$$

We choose the invariant

$$z=\beta(t)\sin y \tag{3.13}$$

as our similarity variable.

By setting $A=A(z)$, $B=B(z)$, we obtain from (3.9)–(3.11) invariant solutions of the form

$$
\begin{aligned}
u&=e^{-s\epsilon(x,t)}[c(x)|\sin y|]^{1/2}[|\beta(t)\cos y+\beta'(t)|^{1/2}A(z)\\
&\quad+|\beta(t)\cos y-\beta'(t)|^{1/2}B(z)],
\end{aligned}
\tag{3.14}
$$

$$
\begin{aligned}
v&=e^{-s\epsilon(x,t)}[[c(x)]^{-1}|\sin y|]^{1/2}[|\beta(t)\cos y\\
&\quad+\beta'(t)|^{1/2}A(z)-|\beta(t)\cos y-\beta'(t)|^{1/2}B(z)],
\end{aligned}
\tag{3.15}
$$

where

$$\epsilon(x,t)=(1/2\nu)\log|\cos y+\beta'(t)\sin y|\quad\text{for }\gamma=-1, \tag{3.16}$$

and

$$\epsilon(x,t)=(1/2\nu)\arctan[\cot y/\beta'(t)]\quad\text{for }\gamma=1. \tag{3.17}$$

The substitution of (3.14) and (3.15) into the system (1.2) leads to a coupled system of first-order linear ordinary differential equations for $A(z)$ and $B(z)$. The form of these ODE's depends on the signs of γ and $\beta(t)\cos y+\beta'(t)$.

If $\gamma=1$, then either

$$\beta(t)\cos y+\beta'(t)>0\quad\text{and}\quad\beta(t)\cos y-\beta'(t)<0$$

or

$$\beta(t)\cos y+\beta'(t)<0\quad\text{and}\quad\beta(t)\cos y-\beta'(t)>0$$

for all x, t.

If $\gamma=-1$, then for any given t, both $\beta(t)\cos y+\beta'(t)$ and $\beta(t)\cos y-\beta'(t)$ change sign once as x varies from $-\infty$ to $+\infty$.

It is convenient to let

$$
\begin{aligned}
&A(z)=(\mathrm{sgn}[\beta(t)\cos y+\beta'(t)])f(z),\\
&B(z)=g(z).
\end{aligned}
\tag{3.18}
$$

Then $\{f(z),g(z)\}$ satisfy the system

$$
\begin{aligned}
&2(z^2-1)\frac{df}{dz}+\left[-\frac{s}{\nu}+\left(2-\frac{s}{\nu}\right)z\right]f\\
&\quad-(1/\nu)|z^2-1|^{1/2}g=0,\\
&2(z^2-1)\frac{dg}{dz}+\left[\frac{s}{\nu}+\left(2-\frac{s}{\nu}\right)z\right]g\\
&\quad+\frac{1}{\nu}\frac{z^2-1}{|z^2-1|^{1/2}}f=0,
\end{aligned}
\tag{3.19}
$$

if $\gamma = -1$, and satisfy the system

$$2(z^2+1)\frac{df}{dz} + \left[2z - \frac{s}{\nu}\right] f - \frac{1}{\nu}\sqrt{z^2+1}\, g = 0,$$
$$2(z^2+1)\frac{dg}{dz} + \left[2z + \frac{s}{\nu}\right] g + \frac{1}{\nu}\sqrt{z^2+1}\, f = 0, \tag{3.20}$$

if $\gamma = 1$.

The invariant solutions for $\gamma = -1$ are not valid for all $t > 0$ since (3.19) has a singular point at $z = 1$. For the rest of this paper we consider solutions of system (1.2) for $\gamma = 1$.

IV. INVARIANT SOLUTIONS OF SYSTEM (1.2) FOR $\gamma = 1$

A. The general solution of (3.20)

Let

$$R = 1/2\nu, \quad \sigma = -s/2\nu. \tag{4.1}$$

Then $f(z)$ satisfies the equation

$$\frac{d^2f}{dz^2} + \frac{3z}{z^2+1}\frac{df}{dz} + \frac{1}{z^2+1}\left[1 + R^2 - \frac{\sigma^2 + \sigma z}{z^2+1}\right] f = 0 \tag{4.2}$$

and

$$g(z) = \frac{\sqrt{z^2+1}}{R}\left[\frac{df}{dz} + \frac{z+\sigma}{z^2+1} f\right]. \tag{4.3}$$

The general solution of (4.2) is

$$\begin{aligned} f(z) = {} & C_1\left(\frac{1-iz}{1+iz}\right)^{i\sigma/2} \\ & \times F\left(1 + iR, 1 - iR; \tfrac{3}{2} - i\sigma; \tfrac{1}{2}(1+iz)\right) \\ & + C_2(1+iz)^{-1/2}(z^2+1)^{i\sigma/2} \\ & \times F\left(\tfrac{1}{2} + i(\sigma+R), \tfrac{1}{2}\right. \\ & \left. + i(\sigma-R); \tfrac{1}{2} + i\sigma; \tfrac{1}{2}(1+iz)\right), \end{aligned} \tag{4.4}$$

where $F(a,b;c;z)$ is the hypergeometric function,[3] C_1 and C_2 are arbitrary constants.

Let

$$\Psi(z) = \log(z + \sqrt{z^2+1}) \tag{4.5}$$

and

$$\tilde{f}(\Psi) = \sqrt{z^2+1}\, f(z).$$

Then (4.2) transforms to

$$\frac{d^2\tilde{f}}{d\Psi^2} + \left[R^2 - \frac{\sigma^2 + \sigma\sinh\Psi}{\cosh^2\Psi}\right]\tilde{f} = 0. \tag{4.6}$$

B. Closed form solutions of (3.20)

Now we construct closed form solutions of (4.2) and (4.3) for various values of σ. From (4.5) and (4.6) we see that for $\sigma = 0$,

$$f = f_0(z;\zeta) = (1/\sqrt{z^2+1})\cos[R\Psi(z) + \zeta] \tag{4.7}$$

solves (4.2) for any real constant ζ. Correspondingly, from (4.3) one gets

$$g = g_0(z;\zeta) = -(1/\sqrt{z^2+1})\sin[R\Psi(z) + \zeta]. \tag{4.8}$$

Now consider the raising and lowering operators

$$L^+(\lambda) = \sqrt{z^2+1}\frac{d}{dz} + \frac{(1-2\lambda)z + i}{2\sqrt{z^2+1}}, \tag{4.9}$$

$$L^-(\lambda) = \sqrt{z^2+1}\frac{d}{dz} + \frac{(1+2\lambda)z - i}{2\sqrt{z^2+1}}. \tag{4.10}$$

One can show that if $f = f_\lambda(z;\zeta)$ solves (4.2) for $\sigma = -i\lambda$, then

$$f = L^+(\lambda) f_\lambda(z;\zeta) \tag{4.11}$$

solves (4.2) for $\sigma = -i(\lambda+1)$, and

$$f = L^-(\lambda) f_\lambda(z;\zeta) \tag{4.12}$$

solves (4.2) for $\sigma = -i(\lambda-1)$.

For $\sigma = -in$, $n = 1,2,\ldots$, recursively we can obtain closed form solutions

$$f = f_n(z;\zeta) = L^+(n-1) f_{n-1}(z;\zeta), \quad n = 1,2,\ldots, \tag{4.13}$$

for (4.2) from $f_0(z;\zeta)$ defined by (4.7).

From (4.3), the corresponding solution is

$$g = g_n(z;\zeta) = \frac{\sqrt{z^2+1}}{R}\left[\frac{df_n(z;\zeta)}{dz} + \frac{z - in}{z^2+1} f_n(z;\zeta)\right]. \tag{4.14}$$

From (4.7)–(4.11), it then follows that

$$L^-(n) f_n(z;\zeta) = -\tfrac{1}{4}[(2n-1)^2 + 4R^2] f_{n-1}(z;\zeta),$$
$$n = 1,2,\ldots\,. \tag{4.15}$$

Using (4.13)–(4.15), one can show that

$$\begin{bmatrix} f_{n+1}(z;\zeta) \\ g_{n+1}(z;\zeta) \end{bmatrix} = \begin{pmatrix} a(n,z) & R \\ -R & \overline{a(n,z)} \end{pmatrix} \begin{bmatrix} f_n(z;\zeta) \\ g_n(z;\zeta) \end{bmatrix}, \tag{4.16}$$

where

$$a(n,z) = -(n + \tfrac{1}{2})[(z - i)/\sqrt{z^2+1}], \quad n = 0,1,2,\ldots,$$

and $\overline{a(n,z)}$ is the complex conjugate of $a(n,z)$.

Let

$$\phi = \operatorname{arccot} z.$$

Then

$$a(n,z) = -(n + \tfrac{1}{2})e^{-i\phi}.$$

In computing $\{f_n(z;\zeta), g_n(z;\zeta)\}$ it is useful to note that

$$\begin{bmatrix} f_n(z;\zeta) \\ g_n(z;\zeta) \end{bmatrix} = \begin{pmatrix} A(n,z,R) & B(n,z,R) \\ -\overline{B(n,z,R)} & \overline{A(n,z,R)} \end{pmatrix} \begin{bmatrix} f_0(z;\zeta) \\ g_0(z;\zeta) \end{bmatrix} \tag{4.17}$$

for functions $A(n,z,R)$ and $B(n,z,R)$ determined from (4.16), $n = 1,2,\ldots$. Further details on $\{f_n(z;\zeta), g_n(z;\zeta)\}$ are given in the Appendix.

Say λ is real. Then $f = f_\lambda(z;\zeta)$ solves (4.2) for $\sigma = -i\lambda$ if and only if the complex conjugate of $f_\lambda(z;\zeta)$, namely $f = \overline{f_\lambda(z;\zeta)}$ solves (4.2) for $\sigma = i\lambda$. Consequently, for $\sigma = in$, $n = 1,2,\ldots$, we have closed form solutions

$$f = f_{-n}(z;\zeta) = \overline{f_n(z;\zeta)},$$
$$g = g_{-n}(z;\zeta) = \overline{g_n(z;\zeta)},$$

for system (3.20).

Moreover if λ is real, using (3.20), one can prove that

$$|f_\lambda(z;\zeta)|^2+|g_\lambda(z;\zeta)|^2 = \frac{\text{const}}{1+z^2}=\frac{|f_\lambda(0;\zeta)|^2+|g_\lambda(0;\zeta)|^2}{1+z^2}. \tag{4.18}$$

Other sequences of solutions are found by the standard technique of using the raising and lowering operators (4.9) and (4.10). Namely we first find functions f which satisfy (4.2) and $L^-(\lambda)f=0$ or $L^+(\lambda)f=0$ for particular values of $\lambda=i\sigma$. For the lowering operator $L^-(\lambda)$, resulting solutions are

$$f=\tilde{f}_0^{\pm}=(1/\sqrt{z^2+1})e^{(i/2)[\arctan z\mp R\log(z^2+1)]} \tag{4.19}$$

for $\lambda=\frac{1}{2}\pm iR$. For the raising operator $L^+(\lambda)$, solutions are

$$f=\hat{f}_0^{\pm}=(1/\sqrt{z^2+1})e^{-(i/2)[\arctan z\pm R\log(z^2+1)]} \tag{4.20}$$

for $\lambda=-\frac{1}{2}\pm iR$. Sequences of solutions $\{\tilde{f}_0^{\pm},\tilde{f}_1^{\pm},\tilde{f}_2^{\pm},\ldots\}$, $\{\hat{f}_0^{\pm},\hat{f}_{-1}^{\pm},\hat{f}_{-2}^{\pm},\ldots\}$, are then obtained as follows:

$$\tilde{f}_{n+1}^{\pm}=L^+(n+\tfrac{1}{2}\pm iR)\tilde{f}_n^{\pm} \tag{4.21}$$

for $\lambda=n+\frac{3}{2}\pm iR$, $n=0,1,2,\ldots$, and

$$\hat{f}_{n-1}^{\pm}=L^-(n-\tfrac{1}{2}\pm iR)\hat{f}_n^{\pm} \tag{4.22}$$

for $\lambda=n-\frac{3}{2}\pm iR$, $n=0,-1,-2,\ldots$.

C. Properties of solutions (3.14) and (3.15)

Since the solutions (3.14) and (3.15) depend on similarity variable z we examine the similarity curves

$$z=\text{const}=\beta(t)\sin y=\beta(t)\sin[\nu\log c(x)], \tag{4.23}$$

where

$$\beta(t)=pe^t-qe^{-t},\quad pq=\tfrac{1}{4}. \tag{4.24}$$

We consider solutions for $t\in(-\infty,\infty)$. Then without loss of generality we can set $p=q=\frac{1}{2}$ by a suitable choice of initial time t, so that

$$\beta(t)=\sinh t. \tag{4.25}$$

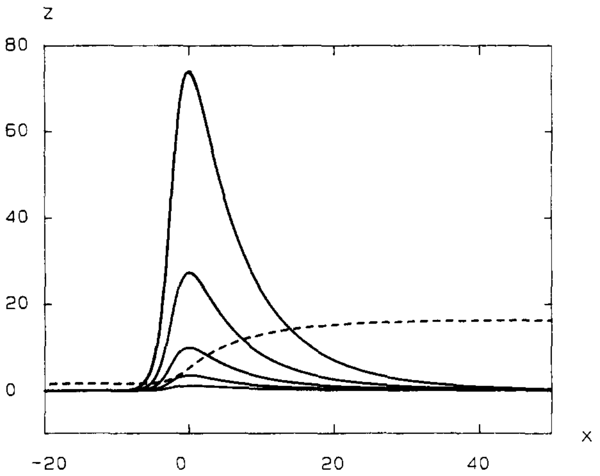

FIG. 4. z as a function of x; z is plotted as a function of x for $\nu=1.4$ and selected values of t: $t=1$ (top), 2, 3, 4, 5 (bottom).

Representative similarity curves are plotted in Fig. 3 for various value of z for $\nu=1.4$. For various values of t, curves

$$z(x,t)=\sinh t\sin[\nu\log c(x)] \tag{4.26}$$

are plotted in Fig. 4 for $\nu=1.4$. Note that $z(0,t)=\sinh t$, $\lim_{x\to\pm\infty}z(x,t)=0$ and hence for fixed t, the range of $z(x,t)$ is $(0,\sinh t]$ if $t>0$ and $[\sinh t,0)$ if $t<0$.

Consider the asymptotic properties of the similarity curves

$$z=\sinh t\sin[\nu\log c(x)]=\text{const}\quad\text{as}\ t\to+\infty.$$

From (2.11) and (2.12), along such curves we have

$$x\sim-(t-\log[2z/\nu C^-(\nu)])\quad\text{if}\ x<0;$$

$$x\sim e^{\pi/\nu}(t-\log[2z/\nu C^+(\nu)])\quad\text{if}\ x>0.$$

Hence as $t\to+\infty$ the similarity curves are asymptotic to the characteristic curves of the wave equation (1.1) or system (1.2). For comparison with the similarity curves of Fig. 3, characteristic curves are plotted in Fig. 5 for $\nu=1.4$.

Next we consider properties of $\{f(z),g(z)\}$. First of all note that $f(z)$ and $g(z)$ are analytic in z. For any σ, as

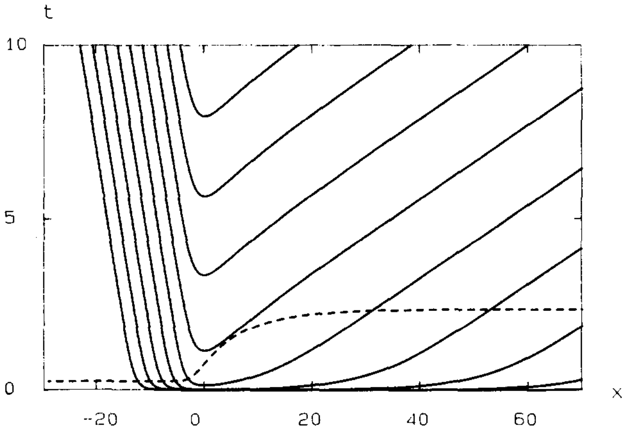

FIG. 3. Similarity curves $z=(\sinh t)\sin(\nu\log c(x))$. Nine similarity curves are plotted for $\nu=1.4$. The corresponding values of z are $z=10^n$ with $n=3$ (top line), 2, 1, 0, -1, -2, -3, -4, -5 (bottom line). The dashed line represents the profile of $c(x)$ for $\nu=1.4$.

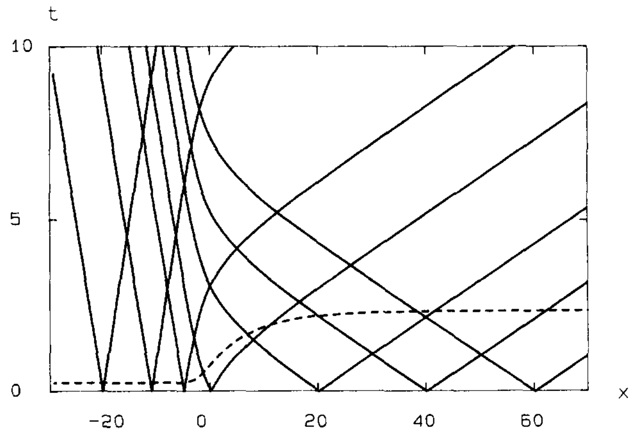

FIG. 5. Characteristic curves, defined by $dx/dt=\pm c(x)$, emanating from the x axis, are plotted for $\nu=1.4$.

$|z| \to \infty$, from (3.20), $\{f(z),g(z)\}$ satisfy

$$z\frac{df}{dz}+f-Rg=0, \quad z\frac{df}{dz}+Rf+g=0. \tag{4.27}$$

Thus

$$\begin{aligned} f(z) &\sim (\mu^{\pm}/z)\cos[R\log|z|+\rho^{\pm}], \\ g(z) &\sim -(\mu^{\pm}/z)\sin[R\log|z|+\rho^{\pm}], \end{aligned} \tag{4.28}$$

as $z \to \pm\infty$ for some constants $\{\mu^+,\mu^-,\rho^+,\rho^-\}$. Thus $\{f(z),g(z)\}$ exhibit oscillatory algebraic decay as $|z| \to \infty$. In the Appendix, μ^+ and ρ^+ are computed for $\{f_n(z;\zeta),g_n(z;\zeta)\}$.

Now consider properties of $\epsilon(x,t)$ defined by Eq. (3.17) with $\beta(t)=\sinh t$, i.e.,

$$\epsilon(x,t)=(1/2\nu)\arctan[\operatorname{sech} t\cot y]. \tag{4.29}$$

The range of $\epsilon(x,t)$ is $(-\pi/4\nu,\pi/4\nu)$ for any t. If $\sigma=-s/2\nu=\sigma_1+i\sigma_2$, then for any $t=t^*$ the number of oscillations with respect to x in a real solution $\{u(x,t^*),v(x,t^*)\}$ due to the factor $e^{-s\epsilon(x,t^*)}$ is the integer n such that

$$n \leqslant \tfrac{1}{2}|\sigma_2| < n+1, \tag{4.30}$$

and for any $x=x^*$ the number of oscillations with respect to t in a real solution $\{u(x^*,t),\ v(x^*,t)\}$ due to the factor $e^{-s\epsilon(x^*,t)}$ is the integer m such that

$$m \leqslant \tfrac{1}{2}|\sigma_2|\cdot|\tfrac{1}{2}-(\nu/\pi)\log c(x^*)| < m+1. \tag{4.31}$$

Furthermore,

$$\epsilon(x,0)=\frac{1}{2\nu}\left[\frac{\pi}{2}-y\right]=\frac{1}{2\nu}\left[\frac{\pi}{2}-\nu\log c(x)\right], \tag{4.32}$$

$$\epsilon(0,t)=0, \tag{4.33}$$

$$\lim_{t\to\pm\infty}\epsilon(x,t)=0, \tag{4.34}$$

and

$$\lim_{x\to\pm\infty}\epsilon(x,t)=\mp(\pi/4\nu). \tag{4.35}$$

At $t=0$,

$$\begin{aligned} u(x,0) &= e^{-s\epsilon(x,0)}[c(x)\sin[\nu\log c(x)]]^{1/2} \\ &\quad\times[f(0)+g(0)], \\ v(x,0) &= e^{-s\epsilon(x,0)}[c(x)]^{-1/2}[\sin[\nu\log c(x)]]^{1/2} \\ &\quad\times[f(0)-g(0)], \end{aligned} \tag{4.36}$$

where $\epsilon(x,0)$ is given by (4.32). In both the real and imaginary parts of (4.36) the number of oscillations with respect to x is the integer n given by (4.30).

At $x=0$,

$$\begin{aligned} u(0,t) &= e^{\pi/4\nu}\sqrt{\cosh t}\,[f(\sinh t)+g(\sinh t)], \\ v(0,t) &= e^{-\pi/4\nu}\sqrt{\cosh t}\,[f(\sinh t)-g(\sinh t)]. \end{aligned} \tag{4.37}$$

Thus $\{u(0,t),\ v(0,t)\}$ are finite in t.

Moreover,

$$\lim_{x\to\pm\infty}u(x,t)=\lim_{x\to\pm\infty}v(x,t)=0. \tag{4.38}$$

Let

$$\Theta(x,t)=R[|t|+\log[\tfrac{1}{2}\sin y]]. \tag{4.39}$$

Then as $t\to+\infty$, x fixed,

$$\begin{aligned} u(x,t) &\sim 2\mu^+[c(x)/\sin y]^{1/2}e^{-t/2} \\ &\quad\times\cos[\Theta(x,t)+\tfrac{1}{2}y+\rho^+], \\ v(x,t) &\sim 2\mu^+[c(x)\sin y]^{-1/2}e^{-t/2} \\ &\quad\times\cos[\Theta(x,t)-\tfrac{1}{2}y+\rho^+]. \end{aligned} \tag{4.40}$$

As $t\to-\infty$, x fixed,

$$\begin{aligned} u(x,t) &\sim 2\mu^-[c(x)/\sin y]^{1/2}e^{t/2} \\ &\quad\times\sin[\Theta(x,t)-\tfrac{1}{2}y+\rho^-], \\ v(x,t) &\sim -2\mu^-[c(x)\sin y]^{-1/2}e^{t/2} \\ &\quad\times\sin[\Theta(x,t)+\tfrac{1}{2}y+\rho^-]. \end{aligned} \tag{4.41}$$

More importantly as $t\to+\infty$ along a similarity curve $z=$ const, one can show that if $x<0$, then

$$\begin{aligned} u(x,t) &= \sqrt{2z}\,e^{\sigma\arctan(1/z)}f(z)[1+(z/\nu)e^{-t}+o(e^{-t})], \\ v(x,t) &= \sqrt{2z}\,e^{\sigma\arctan(1/z)}f(z)[1-(z/\nu)e^{-t}+o(e^{-t})]; \end{aligned} \tag{4.42}$$

if $x>0$, then

$$\begin{aligned} u(x,t) &= \sqrt{2z}\,e^{\pi/2\nu}e^{-\sigma\arctan(1/z)} \\ &\quad\times g(z)[1-(z/\nu)e^{-t}+o(e^{-t})], \\ v(x,t) &= -\sqrt{2z}\,e^{-\pi/2\nu}e^{-\sigma\arctan(1/z)} \\ &\quad\times g(z)[1+(z/\nu)e^{-t}+o(e^{-t})]. \end{aligned} \tag{4.43}$$

V. SUPERPOSITION OF INVARIANT SOLUTIONS; SOLUTION OF THE INITIAL VALUE PROBLEM

By superposing invariant solutions, general initial value problems (IVP's) of the form

$$u(x,0)=U(x), \quad v(x,0)=V(x), \quad -\infty<x<\infty, \tag{5.1}$$

for system (1.2), and

$$u(x,0)=U(x), \quad u_t(x,0)=W(x), \quad -\infty<x<\infty, \tag{5.2}$$

for Eq. (1.1), can be solved. Solutions $u(x,t)$ of (1.1) and (1.2) are identical if

$$W(x)=c^2(x)V'(x). \tag{5.3}$$

For $\sigma=-2mi$, i.e., $s=4\nu mi$, $m=0,\pm1,\pm2,\ldots$, consider invariant solutions (3.14) and (3.15) of system (1.2) $u=u_m(x,t;\zeta_{2m})$, $v=v_m(x,t;\zeta_{2m})$,

$$\begin{aligned} &u_m(x,t;\zeta_{2m}) \\ &\quad=\exp(-i2m\arctan[\cot y\operatorname{sech} t])\cdot[c(x)\sin y]^{1/2} \\ &\qquad\times\{[\cosh t+\sinh t\cos y]^{1/2}f_{2m}(z;\zeta_{2m}) \\ &\qquad+[\cosh t-\sinh t\cos y]^{1/2}g_{2m}(z;\zeta_{2m})\}, \end{aligned} \tag{5.4}$$

$$\begin{aligned} v_m(x,t;\zeta_{2m}) &= \exp(-i2m\arctan[\cot y\operatorname{sech} t])\cdot\left[\frac{\sin y}{c(x)}\right]^{1/2} \\ &\quad\times\{[\cosh t+\sinh t\cos y]^{1/2}f_{2m}(z;\zeta_{2m}) \\ &\quad-[\cosh t-\sinh t\cos y]^{1/2}g_{2m}(z;\zeta_{2m})\}, \end{aligned} \tag{5.5}$$

where $\{f_{2m}(z;\zeta), g_{2m}(z;\zeta)\}$ are defined by (4.7), (4.8), and (4.16).

At $t=0$,

$$u_m(x,0;\zeta_{2m}) = (-1)^m [c(x)\sin y]^{1/2} [f_{2m}(0;\zeta_{2m}) + g_{2m}(0;\zeta_{2m})] e^{i2my}, \tag{5.6}$$

$$v_m(x,0;\zeta_{2m}) = (-1)^m [\sin y / c(x)]^{1/2} [f_{2m}(0;\zeta_{2m}) - g_{2m}(0;\zeta_{2m})] e^{i2my}. \tag{5.7}$$

For solving an initial value problem it is necessary that $\zeta_{-2m} = \zeta_{2m}$. Note that $0<2y<2\pi$. We let a superposition of invariant solutions,

$$\begin{aligned} u(x,t) &= \sum_{m=-\infty}^{\infty} A_m u_m(x,t;\zeta_{2m}),\\ v(x,t) &= \sum_{m=-\infty}^{\infty} A_m v_m(x,t;\zeta_{2m}), \end{aligned} \tag{5.8}$$

represent the solution of the initial value problem (5.1) for system (1.2). The constants $\{A_m, \zeta_{2m}\}$ are to be determined. In practice we determine $\{A_m \cos \zeta_{2m}, A_m \sin \zeta_{2m}\}$ due to the form of (5.8). Clearly $A_{-m} = \overline{A}_m$ since $u(x,t)$ and $v(x,t)$ are real.

The initial condition (5.1) and (5.6)–(5.8) lead to the following Fourier series representations:

$$\begin{aligned} U(x)[c(x)\sin y]^{-1/2} &= \sum_{m=-\infty}^{\infty} B_m e^{i2my},\\ V(x)\left[\frac{c(x)}{\sin y}\right]^{1/2} &= \sum_{m=-\infty}^{\infty} C_m e^{i2my}, \end{aligned} \tag{5.9}$$

where

$$\begin{aligned} B_m &= (-1)^m [f_{2m}(0;\zeta_{2m}) + g_{2m}(0;\zeta_{2m})] A_m,\\ C_m &= (-1)^m [f_{2m}(0;\zeta_m) - g_{2m}(0;\zeta_{2m})] A_m,\\ &\qquad m = 0, \pm 1, \pm 2,..., \end{aligned} \tag{5.10}$$

$$\begin{aligned} B_m &= \frac{1}{\pi}\int_0^\pi e^{-i2my} U(x(y)) e^{-y/2\nu} [\sin y]^{-1/2}\, dy,\\ C_m &= \frac{1}{\pi}\int_0^\pi e^{-i2my} V(x(y)) e^{y/2\nu} [\sin y]^{-1/2}\, dy, \end{aligned} \tag{5.11}$$

where x and y are related in a 1:1 manner by $y = \nu \log c(x)$. This completes the solution of the IVP (5.1) of system (1.2). The convergence properties of the Fourier series (5.9) depend on the nature of the functions $U(x)[c(x)\sin y]^{-1/2}$, $V(x)[c(x)/\sin y]^{1/2}$. If $\lim_{x\to\pm\infty} U(x) = \lim_{x\to+\infty} V(x) = 0$, $U(x)$, $V(x)$ bounded on $(-\infty,\infty)$ then the series (5.9) converge in the mean.

See the Appendix for general expressions for $\{f_{2m}(0;\zeta) \pm g_{2m}(0;\zeta)\}$ and discussion of the algorithm to compute (5.8).

Now we give the algorithm to find the Green's functions $(G_i(x,\xi,t), K_i(x,\xi,t))$, $i=1,2$, for the initial value problem (5.1). Here $(u,v) = (G_i(x,\xi,t), K_i(x,\xi,t))$, $i=1,2$, satisfies (1.2), and

$$\begin{aligned} &G_1(x,\xi,0) = \delta(x-\xi), \quad K_1(x,\xi,0) = 0;\\ &G_2(x,\xi,0) = 0, \quad K_2(x,\xi,0) = \delta(x-\xi). \end{aligned} \tag{5.12}$$

In terms of these Green's functions, the solution of the IVP (5.1) for system (1.2) may be formally represented as

$$\begin{aligned} u &= \int_{-\infty}^{\infty} [G_1(x,\xi,t)U(\xi) + G_2(x,\xi,t)V(\xi)]\,d\xi,\\ v &= \int_{-\infty}^{\infty} [K_1(x,\xi,t)U(\xi) + K_2(x,\xi,t)V(\xi)]\,d\xi. \end{aligned} \tag{5.13}$$

In computing the coefficients for (G_i, K_i) we set $C_m = C^i_m$, $B_m = B^i_m$, $A_m = A^i_m$, $\zeta_{2m} = \zeta^i_{2m}$, $i=1,2$. Then (5.11) gives

$$\begin{aligned} C^1_m &= 0,\\ B^1_m &= \frac{1}{\pi}[c(\xi)]^{-3/2}[\sin[\nu \log c(\xi)]]^{1/2} e^{-i2m\nu \log c(\xi)},\\ C^2_m &= c(\xi) B^1_m,\\ B^2_m &= 0. \end{aligned} \tag{5.14}$$

Now from (5.10), (5.14), (5.8), (5.4), and (5.5) it follows that $\{G_1, K_1, G_2, K_2\}$ are of the form

$$\begin{aligned} G_1(x,\xi,t) &= [c(\xi)]^{-3/2}[\sin[\nu \log c(\xi)]^{1/2}]\\ &\quad \times \sum_{m=-\infty}^{\infty} a^1_m e^{-i2m\nu \log c(\xi)} U^1_m(x,t),\\ K_1(x,\xi,t) &= [c(\xi)]^{-3/2}[\sin[\nu \log c(\xi)]]^{1/2}\\ &\quad \times \sum_{m=-\infty}^{\infty} b^1_m e^{-i2m\nu \log c(\xi)} V^1_m(x,t), \end{aligned} \tag{5.15}$$

$$\begin{aligned} G_2(x,\xi,t) &= [c(\xi)]^{-1/2}[\sin[\nu \log c(\xi)]]^{1/2}\\ &\quad \times \sum_{m=-\infty}^{\infty} a^2_m e^{-i2m\nu \log c(\xi)} U^2_m(x,t),\\ K_2(x,\xi,t) &= [c(\xi)]^{-1/2}[\sin[\nu \log c(\xi)]]^{1/2}\\ &\quad \times \sum_{m=-\infty}^{\infty} b^2_m e^{-i2m\nu \log c(\xi)} V^2_m(x,t), \end{aligned} \tag{5.15}$$

where the constants $\{a^i_m, b^i_m\}$ and the functions $\{U^i_m(x,t), V^i_m(x,t)\}$, $i=1,2$, are independent of ξ.

Now consider (5.11) for hump functions (unimodal functions)

$$U(x) = (\sin y)^{n+1/2} e^{(1/2)\alpha y}, \quad V(x) = 0, \tag{5.16}$$

where $n = 0,1,2,...$, and α is an arbitrary real constant.

Then $\lim_{x\to\pm\infty} U(x) = 0$, and $U(x)$ has precisely one extremum (a maximum) located at $y = y^\dagger$, $0 < y^\dagger < \pi$, where

$$y^\dagger = \operatorname{arccot}(-\alpha/(2n+1)). \tag{5.17}$$

Let

$$\kappa = \alpha/(2n+1), \tag{5.18}$$

$$U(x;\kappa,n) = [\sin y e^{\kappa y} / \sin y^\dagger e^{\kappa y^\dagger}]^{n+1/2}, \quad n = 0,1,2,... . \tag{5.19}$$

For each n, the hump function $U(x) = U(x;\kappa,n)$ has amplitude 1 with its maximum located at $y = y^\dagger = \operatorname{arccot}(-\kappa)$.

If $y^\dagger$ is fixed and n increases, from (5.19) it follows that the hump sharpens. It sharpens to a spike as $n\to\infty$. Three profiles of $U(x)$ are plotted in Figs. 6(a) and 6(b) for $n=0$ and $n=10$, respectively, with $\nu = 1.4$.

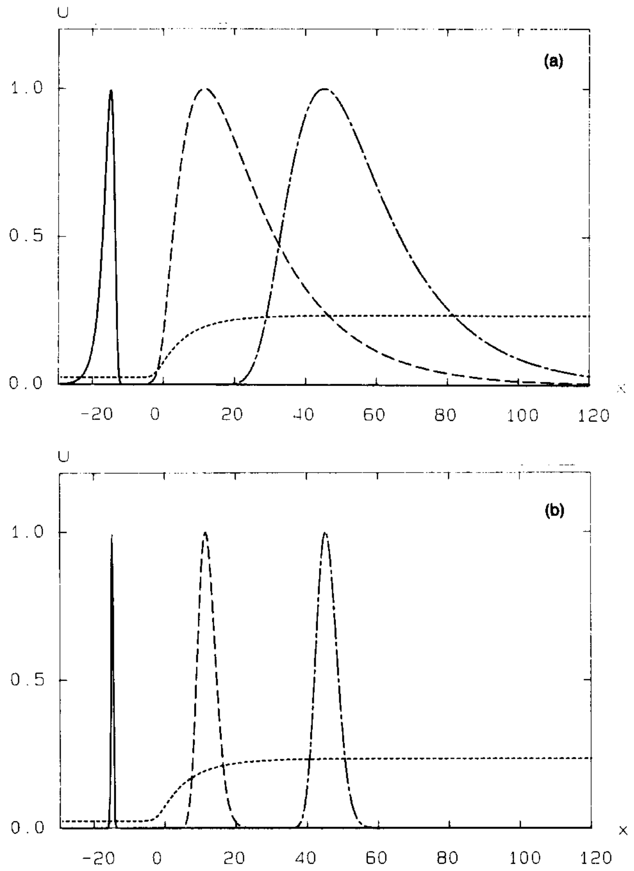

FIG. 6. (a) Hump function $U(x;\kappa,0)$; (b) Hump function $U(x;\kappa,10)$. Three hump functions are plotted for $n=0$ [Fig. (a)] and $n=10$ [Fig. (b)]. In both cases the locations of the peaks are at $x=-15, 11.25$, and 45 and the corresponding values of κ are about -2.25×10^5, 3.73, and 1.62×10^2, respectively. The value of ν is 1.4.

Let

$$A(\kappa,n)=[\sin y^\dagger e^{\kappa y^\dagger}]^{-(n+1/2)}. \tag{5.20}$$

Corresponding to $U(x;\kappa,n)$,

$$B_m=B_m(\kappa,n)=\frac{A(\kappa,n)}{\pi}\int_0^\pi e^{-i2my}e^{b(\kappa,n)y}\sin^n y\,dy,$$

which integrates to

$$B_m=n!\,[A(\kappa,n)/a\pi]$$
$$\times\frac{(e^{b\pi}-1)}{(a^2+n^2)(a^2+(n-2)^2)\cdots(a^2+2^2)}$$
$$\text{if } n=2N,\quad N=1,2,...,$$

$$B_m=n!\,[A(\kappa,n)/\pi]$$
$$\times\frac{(e^{b\pi}+1)}{(a^2+n^2)(a^2+(n-2)^2)\cdots(a^2+1^2)}$$
$$\text{if } n=2N-1,\quad N=1,2,..., \tag{5.21}$$

with

$$b=b(\kappa,n)=\tfrac{1}{2}[(2n+1)\kappa-1/\nu], \tag{5.22}$$

and

$$a=a(\kappa,m,n)=b(\kappa,n)-2mi. \tag{5.23}$$

One can show that

$$B_m(\kappa,2N)=\frac{(2N)!\,A(\kappa,2N)}{\pi}(e^{b\pi}-1)\frac{b+2mi}{b^2+4m^2}$$
$$\times\prod_{k=1}^{N}\frac{b^2+4k^2-4m^2+4bmi}{(b^2+4k^2-4m^2)^2+16b^2m^2},$$
$$B_m(\kappa,2N-1)$$
$$=\frac{(2N-1)!\,A(\kappa,2N-1)}{\pi}(e^{b\pi}+1) \tag{5.24}$$
$$\times\prod_{k=1}^{N}\frac{b^2+(2k-1)^2-4m^2+4bmi}{(b^2+(2k-1)^2-4m^2)^2+16b^2m^2},$$
$$N=1,2,\dots.$$

If $n=0$,

$$B_m(\kappa,0)=(2/\pi)A(\kappa,0)[e^{(1/2)(\kappa-1/\nu)\pi}-1]$$
$$\times\left[\frac{(\kappa-1/\nu)+4mi}{(\kappa-1/\nu)^2+16m^2}\right]. \tag{5.25}$$

ACKNOWLEDGMENT

This work was supported in part by the Natural Sciences and Engineering Council of Canada under Grant No. A7157.

APPENDIX

1. Computation of $\{f_{2m}(0;\zeta)\pm g_{2m}(0;\zeta)\}$

From (4.16), it follows that

$$\begin{bmatrix}f_{n+1}(0;\zeta)\\ g_{n+1}(0;\zeta)\end{bmatrix}=\begin{pmatrix}i(n+\frac{1}{2}) & R\\ -R & -i(n+\frac{1}{2})\end{pmatrix}\begin{bmatrix}f_n(0;\zeta)\\ g_n(0;\zeta)\end{bmatrix},$$
$$n=0,1,2,\dots. \tag{A1}$$

Hence

$$\begin{bmatrix}f_{n+1}(0;\zeta)+g_{n+1}(0;\zeta)\\ f_{n+1}(0;\zeta)-g_{n+1}(0;\zeta)\end{bmatrix}=\begin{pmatrix}0 & i(n+\frac{1}{2})-R\\ i(n+\frac{1}{2})+R & 0\end{pmatrix}\begin{bmatrix}f_n(0;\zeta)+g_n(0;\zeta)\\ f_n(0;\zeta)-g_n(0;\zeta)\end{bmatrix},$$
$$n=0,1,2,\dots. \tag{A2}$$

It follows that

$$\begin{bmatrix}f_{n+2}(0;\zeta)+g_{n+2}(0;\zeta)\\ f_{n+2}(0;\zeta)-g_{n+2}(0;\zeta)\end{bmatrix}=-\begin{pmatrix}(n+\frac{3}{2})(n+\frac{1}{2})+R^2-iR & 0\\ 0 & (n+\frac{3}{2})(n+\frac{1}{2})+R^2+iR\end{pmatrix}\begin{bmatrix}f_n(0;\zeta)+g_n(0;\zeta)\\ f_n(0;\zeta)-g_n(0;\zeta)\end{bmatrix},$$
$$n=0,1,2,\dots. \tag{A3}$$

Let

$$\alpha_m = \arctan\frac{4R}{(4m-1)(4m-3)+4R^2}, \qquad \text{(A4)}$$

$$s_m = ([(4m-1)(4m-3)+4R^2]^2+16R^2)^{1/2}, \qquad \text{(A5)}$$

and

$$\Theta_m = \alpha_1+\alpha_2+\cdots+\alpha_m, \quad m = 1,2,\ldots. \qquad \text{(A6)}$$

Then

$$f_{2m}(0;\zeta) \pm g_{2m}(0;\zeta) = [(-1)^m/4^m](s_1 s_2\cdots s_m)e^{\mp i\Theta_m}[\cos\zeta \mp \sin\zeta], \quad m = 1,2,\ldots. \qquad \text{(A7)}$$

Note that

$$\begin{aligned} f_{2m}(0;\zeta) &= [(-1)^m/4^m](s_1 s_2\cdots s_m) \\ &\quad\times[\cos\Theta_m \cos\zeta + i\sin\Theta_m \sin\zeta], \\ g_{2m}(0;\zeta) &= [(-1)^{m+1}/4^m](s_1 s_2\cdots s_m) \\ &\quad\times[\cos\Theta_m \sin\zeta + i\sin\Theta_m \cos\zeta], \\ m &= 1,2,\ldots. \end{aligned} \qquad \text{(A8)}$$

Thus [cf. (4.18)]

$$(|f_{2m}(0;\zeta)|^2+|g_{2m}(0;\zeta)|^2)^{1/2} = s_1 s_2\cdots s_m/4^m, \quad m = 1,2,\ldots. \qquad \text{(A9)}$$

2. Computation of $\{f_n(z;\zeta), g_n(z;\zeta)\}$

Consider (4.16). The matrix

$$M(n,z,R) \equiv \frac{1}{\sqrt{(n+\frac{1}{2})^2+R^2}}\begin{pmatrix} a(n,z) & R \\ -R & \overline{a(n,z)} \end{pmatrix} \qquad \text{(A10)}$$

is a unitary matrix.

Let

$$\beta_n = \arctan[2R/(2n+1)], \quad n = 0,1,2,\ldots, \qquad \text{(A11)}$$

and

$$f = \mathrm{Re} f + i\,\mathrm{Im} f.$$

Then

$$\begin{bmatrix} \mathrm{Re} f_{n+1}(z;\zeta) \\ \mathrm{Im} f_{n+1}(z;\zeta) \\ \mathrm{Re}\, g_{n+1}(z;\zeta) \\ \mathrm{Im}\, g_{n+1}(z;\zeta) \end{bmatrix} = -\sqrt{(n+\tfrac{1}{2})^2+R^2}\,N(\beta_m,\phi)\begin{bmatrix} \mathrm{Re} f_n(z;\zeta) \\ \mathrm{Im} f_n(z;\zeta) \\ \mathrm{Re}\, g_n(z;\zeta) \\ \mathrm{Im}\, g_n(z;\zeta) \end{bmatrix}, \qquad \text{(A12)}$$

where $\phi = \operatorname{arccot} z$, and the 4×4 orthogonal matrix

$$N(\beta_n,\phi) \equiv \begin{pmatrix} \cos\beta_n\cos\phi & \cos\beta_n\sin\phi & -\sin\beta_n & 0 \\ -\cos\beta_n\sin\phi & \cos\beta_n\cos\phi & 0 & -\sin\beta_n \\ \sin\beta_n & 0 & \cos\beta_n\cos\phi & -\cos\beta_n\sin\phi \\ 0 & \sin\beta_n & \cos\beta_n\sin\phi & \cos\beta_n\cos\phi \end{pmatrix}, \quad n = 0,1,2,\ldots, \qquad \text{(A13)}$$

and

$$\begin{bmatrix} \mathrm{Re} f_0(z;\zeta) \\ \mathrm{Im} f_0(z;\zeta) \\ \mathrm{Re}\, g_0(z;\zeta) \\ \mathrm{Im}\, g_0(z;\zeta) \end{bmatrix} = \frac{1}{\sqrt{z^2+1}}\begin{bmatrix} \cos[R\log(z+\sqrt{z^2+1})+\zeta] \\ 0 \\ -\sin[R\log(z+\sqrt{z^2+1})+\zeta] \\ 0 \end{bmatrix}. \qquad \text{(A14)}$$

3. Asymptotic properties of $\{f_n(z;\zeta), g_n(z;\zeta)\}$

As $z\to+\infty$, from (4.16),

$$\begin{bmatrix} f_{n+1}(z;\zeta) \\ g_{n+1}(z;\zeta) \end{bmatrix} \sim -\sqrt{(n+\tfrac{1}{2})^2+R^2} \times\begin{pmatrix} \cos\beta_n & -\sin\beta_n \\ \sin\beta_n & \cos\beta_n \end{pmatrix}\begin{bmatrix} f_n(z;\zeta) \\ g_n(z;\zeta) \end{bmatrix}, \quad n = 0,1,2,\ldots. \qquad \text{(A15)}$$

From (A15) and an analysis of the error in (A15), one can show that as $z\to+\infty$ [cf. (4.28)],

$$\begin{aligned} f_n(z;\zeta) &= (\mu_n^+/z)[\cos[R\log z+\rho_n^+]] \\ &\quad\times[1+O(1/z)], \\ g_n(z;\zeta) &= -(\mu_n^+/z)\sin[[R\log z+\rho_n^+]] \\ &\quad\times[1+O(1/z)], \end{aligned} \qquad \text{(A16)}$$

where

$$\mu_n^+ = (-1)^n([(n-\tfrac{1}{2})^2+R^2][(n-\tfrac{3}{2})^2+R^2] \times\cdots\times[(\tfrac{1}{2})^2+R^2])^{1/2}, \qquad \text{(A17)}$$

$$\rho_n^+ = R\log 2 + \zeta - \sum_{k=0}^{n-1} \beta_k, \quad n = 1,2,\ldots; \qquad \text{(A18)}$$

$\{\beta_k\}$ defined by (A11).

4. Discussion of the algorithm to compute (5.8)

For $n = 2m$, consider the matrix defined by (4.17), namely

$$P_{2m}(z,R) = \begin{pmatrix} A(2m,z,R) & B(2m,z,R) \\ -\overline{B(2m,z,R)} & \overline{A(2m,z,R)} \end{pmatrix}. \qquad \text{(A19)}$$

Then

$$\begin{bmatrix} f_{2m}(z;\zeta_{2m}) \\ g_{2m}(z;\zeta_{2m}) \end{bmatrix} = P_{2m}(z,R) \begin{bmatrix} f_0(z;\zeta_{2m}) \\ g_0(z;\zeta_{2m}) \end{bmatrix}. \qquad \text{(A20)}$$

Note that the matrix $P_{2m}(z,R)$ is independent of ζ_{2m}. From (4.7) and (4.8),

$$\begin{bmatrix} f_0(z;\zeta_{2m}) \\ g_0(z;\zeta_{2m}) \end{bmatrix} = \frac{1}{\sqrt{z^2+1}} \begin{pmatrix} \cos R\Psi(z) & \sin R\Psi(z) \\ -\sin R\Psi(z) & \cos R\Psi(z) \end{pmatrix} \times \begin{bmatrix} \cos\zeta_{2m} \\ -\sin\zeta_{2m} \end{bmatrix}. \qquad \text{(A21)}$$

Now multiply both sides of (A20) by A_m and set $z = 0$. Then

$$A_m \begin{bmatrix} f_{2m}(0;\zeta_{2m}) \\ g_{2m}(0;\zeta_{2m}) \end{bmatrix} = A_m P_{2m}(0,R) \begin{bmatrix} \cos\zeta_{2m} \\ -\sin\zeta_{2m} \end{bmatrix}, \qquad \text{(A22)}$$

where (A7) gives

$$P_{2m}(0,R) = \frac{(-1)^m s_1 s_2 \cdots s_m}{2\cdot 4^m} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \times \begin{pmatrix} e^{-i\Theta_m} & 0 \\ 0 & e^{i\Theta_m} \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}. \qquad \text{(A23)}$$

Thus from (A22)

$$A_m \begin{bmatrix} \cos\zeta_{2m} \\ -\sin\zeta_{2m} \end{bmatrix} = [P_{2m}(0,R)]^{-1} A_m \begin{bmatrix} f_{2m}(0;\zeta_{2m}) \\ g_{2m}(0;\zeta_{2m}) \end{bmatrix}, \qquad \text{(A24)}$$

with

$$[P_{2m}(0,R)]^{-1} = \frac{(-1)^m 4^m}{2 s_1 s_2 \cdots s_m} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \times \begin{pmatrix} e^{i\Theta_m} & 0 \\ 0 & e^{-i\Theta_m} \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}. \qquad \text{(A25)}$$

From the initial condition (5.10),

$$A_m \begin{bmatrix} f_{2m}(0;\zeta_{2m}) \\ g_{2m}(0;\zeta_{2m}) \end{bmatrix} = \frac{(-1)^m}{2} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{bmatrix} B_m \\ C_m \end{bmatrix}. \qquad \text{(A26)}$$

Hence in the superposition (5.8),

$$A_m \begin{bmatrix} f_{2m}(z;\zeta_{2m}) \\ g_{2m}(z;\zeta_{2m}) \end{bmatrix} = \frac{4^m}{2 s_1 s_2 \cdots s_m \sqrt{z^2+1}} \times P_{2m}(z,R) \begin{pmatrix} \cos R\Psi(z) & \sin R\Psi(z) \\ -\sin R\Psi(z) & \cos R\Psi(z) \end{pmatrix} \times \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} e^{i\Theta_m} & 0 \\ 0 & e^{-i\Theta_m} \end{pmatrix} \begin{bmatrix} B_m \\ C_m \end{bmatrix}. \qquad \text{(A27)}$$

Note that explicit computations of $\{A_m, \zeta_{2m}\}$ are not required. Thus the problem of determining $\{A_m f_{2m}(z;\zeta_{2m}), A_m g_{2m}(z;\zeta_{2m})\}$ has been reduced to the computation of $P_{2m}(z,R)$.

Algebraically, $P_{2m}(z,R)$ is determined by using the recursive relation (4.16) or its real version (A12)–(A14). Next we give a nonrecursive procedure for finding $P_{2m}(z,R)$ based on a numerical solution of an initial value problem for a system of ordinary differential equations.

Let

$$\begin{bmatrix} f \\ g \end{bmatrix} = \begin{bmatrix} F_{2m}(z) \\ G_{2m}(z) \end{bmatrix} \qquad \text{(A28)}$$

solve the system corresponding to (3.20) and (4.1) for $\sigma = -2mi$, namely

$$\begin{aligned} &(z^2+1)\frac{df}{dz} + (z - 2mi)f - R\sqrt{z^2+1}\,g = 0, \\ &(z^2+1)\frac{dg}{dz} + (z + 2mi)g + R\sqrt{z^2+1}\,f = 0, \end{aligned} \qquad \text{(A29)}$$

with initial condition

$$\begin{bmatrix} f(0) \\ g(0) \end{bmatrix} = \frac{(-1)^m}{2} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{bmatrix} b_m \\ c_m \end{bmatrix} \qquad \text{(A30)}$$

for any nontrivial choice of constants $\{b_m, c_m\}$, $m = 1,2,\ldots$. Then $\left[\begin{smallmatrix} F_{2m}(z) \\ G_{2m}(z) \end{smallmatrix}\right]$ equals the right-hand side of (A27) with $B_m = b_m$, $C_m = c_m$. Here $P_{2m}(z,R)$ is determined in terms of $\{F_{2m}(z), G_{2m}(z)\}$,

$$P_{2m}(z,R) = \frac{s_1 s_2 \cdots s_m}{4^m[|B_m|^2 + |C_m|^2]} \cdot \sqrt{z^2+1} \begin{pmatrix} F_{2m}(z) & -\overline{G_{2m}(z)} \\ G_{2m}(z) & \overline{F_{2m}(z)} \end{pmatrix} \times \begin{pmatrix} \overline{B}_m & \overline{C}_m \\ C_m & -B_m \end{pmatrix} \begin{pmatrix} e^{-i\Theta_m} & 0 \\ 0 & e^{+i\Theta_m} \end{pmatrix} \begin{pmatrix} \cos R\Psi(z) & \sin R\Psi(z) \\ -\sin R\Psi(z) & \cos R\Psi(z) \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}. \qquad \text{(A31)}$$

Note that the matrix $P_{2m}(z,R)$ is independent of the choice of $\{b_m, c_m\}$. If $b_m = 1$, $c_m = 0$, then (A31) becomes

$$P_{2m}(z,R) = \frac{s_1 s_2 \cdots s_m}{4^m} \sqrt{2(z^2+1)} \begin{pmatrix} F_{2m}(z) & -\overline{G_{2m}(z)} \\ G_{2m}(z) & \overline{F_{2m}(z)} \end{pmatrix} \begin{pmatrix} e^{-i\Theta_m}\cos(R\Psi(z) - \pi/4) & -e^{-i\Theta_m}\sin(R\Psi(z) - \pi/4) \\ e^{i\Theta_m}\sin(R\Psi(z) - \pi/4) & e^{i\Theta_m}\cos(R\Psi(z) - \pi/4) \end{pmatrix}. \qquad \text{(A32)}$$

Any numerical procedure such as Runge–Kutta can be used to find $\{F_{2m}(z), G_{2m}(z)\}$, $m = 1,2,\ldots$.

The following asymptotic expression is useful for computing $\{A_m f_{2m}(z;\zeta_{2m}), A_m g_{2m}(z;\zeta_{2m})\}$ for large z: using (A11), (A16)–(A18), (A24)–(A26), one can show that as $z \to +\infty$,

$$A_m \begin{bmatrix} f_{2m}(z;\zeta_{2m}) \\ g_{2m}(z;\zeta_{2m}) \end{bmatrix} = \frac{\sqrt{2}}{z}\left[1 + O\left(\frac{1}{z}\right)\right] \frac{4^m \mu_{2m}^+}{2s_1 s_2 \cdots s_m} \times \begin{pmatrix} e^{i\Theta_m}\cos\omega_m(z) & -e^{-i\Theta_m}\sin\omega_m(z) \\ -e^{i\Theta_m}\sin\omega_m(z) & -e^{-i\Theta_m}\cos\omega_m(z) \end{pmatrix} \times \begin{bmatrix} B_m \\ C_m \end{bmatrix}, \tag{A33}$$

where

$$\omega_m(z) = R\log 2z - \frac{\pi}{4} - \sum_{k=0}^{2m-1} \beta_k, \tag{A34}$$

and $\{\Theta_m, s_m\}$ are given by (A4)–(A6).

[1] G. Bluman and S. Kumei, J. Math. Phys. **28**, 307 (1987).

[2] B. Seymour and E. Varley, Stud. Appl. Math. **76**, 1 (1987).

[3] M. Abramowitz and I. Stegun, *Handbook of Mathematical Functions* (Dover, New York, 1965), Chap. 15.

Splitting methods and invariant imbedding for time-independent wave propagation in focusing media and wave guides

J. W. Evans
Ames Laboratory, Applied Mathematical Sciences, Iowa State University, Ames, Iowa 50011

(Received 30 July 1986; accepted for publication 16 September 1987)

For time-independent wave propagation in focusing media or wave guides, backscattering and coupling between propagation modes are caused by deterministic or random variations of the refractive index in the distinguished (x) direction of propagation. Various splittings of the wave field into forward and backward traveling components, which lead to coupled equations involving abstract operator coefficients, are presented. Choosing a natural explicit representation for these operators immediately yields a coupled mode form of these equations. The splitting procedure also leads naturally to abstract transmission and reflection operators for slabs of finite thickness $(a \leqslant x \leqslant b)$, and abstract invariant imbedding equations satisfied by these. The coupled mode form of these equations, together with such features as reciprocity (associated with an underlying symplectic structure) are also discussed. The example of a square law medium is used to illustrate some of these concepts.

I. INTRODUCTION

Here we consider only time-independent scalar wave propagation described by the $d \geqslant 2$ dimensional Helmholtz equation. We assume that there is a distinguished direction of propagation chosen as the x direction in a Cartesian coordinate system $(x_1, x_2, x_3, \ldots)$ where $x_1 = x$, $(x_2, x_3, \ldots) = \mathbf{x}_\perp$. The Helmholtz equation is thus written naturally as

$$\psi_{xx} + S\psi = 0 \text{ or } \frac{d}{dx}\begin{pmatrix}\psi \\ \psi_x\end{pmatrix} = \begin{pmatrix}0 & 1 \\ -S & 0\end{pmatrix}\begin{pmatrix}\psi \\ \psi_x\end{pmatrix}, \qquad (1.1)$$

where $S = \Delta_\perp + k^2(\mathbf{x})$, and suitable boundary conditions are imposed on ψ if the range of $\mathbf{x}_\perp$ is restricted. Here $\Delta_\perp = \partial^2/\partial \mathbf{x}_\perp{}^2$ is the transverse Laplacian, $k(\mathbf{x}) = kn(\mathbf{x})$, where $n(\mathbf{x})$ is the refractive index, and $k > 0$ is arbitrary. We shall regard $\{S \equiv S(x)\}$ (implicitly including any appropriate boundary conditions) as a generally noncommutative family of unbounded self-adjoint operators on $L^2(\mathbf{x}_\perp)$.

Our treatment of the Helmholtz equation (1.1) is based on a splitting of ψ into right (x increasing) ψ^+, and left (x decreasing) ψ^-, traveling components. This decomposition is achieved in terms of a splitting operator P as

$$\begin{pmatrix}\psi^+ \\ \psi^-\end{pmatrix} = P\begin{pmatrix}\psi \\ \psi_x\end{pmatrix}, \text{ with } P = P(x) = \frac{1}{2}\begin{pmatrix}1 & -iT^{-1/2} \\ 1 & +iT^{-1/2}\end{pmatrix}, \qquad (1.2)$$

i.e., $\psi^\pm = \frac{1}{2}(\psi \mp iT^{-1/2}\psi_x)$, so $\psi \equiv \psi^- + \psi^+$. Suitable choices of the operators $T \equiv T(x)$, on $L^2(\mathbf{x}_\perp)$ are discussed below (cf. Refs. 1–6). Formal manipulation of (1.1) now yields (cf. Ref. 4)

$$\frac{d}{dx}\begin{pmatrix}\psi^+ \\ \psi^-\end{pmatrix} = \mathsf{A}(x)\begin{pmatrix}\psi^+ \\ \psi^-\end{pmatrix},$$

$$\mathsf{A}(x) = P\begin{pmatrix}0 & 1 \\ -S & 0\end{pmatrix}P^{-1} + \left(\frac{d}{dx}P\right)P^{-1}, \qquad (1.3\text{a})$$

or

$$\frac{d}{dx}\psi^\pm = \pm\frac{1}{2}iT^{-1/2}[(S+T)\psi^\pm + (S-T)\psi^\mp]$$

$$+\frac{1}{2}(T^{-1/2})_x T^{1/2}(\psi^\pm - \psi^\mp). \qquad (1.3\text{b})$$

To motivate (1.2) and (1.3), we note that the choice $T \equiv S$ diagonalizes

$$P\begin{pmatrix}0 & 1 \\ -S & 0\end{pmatrix}P^{-1},$$

thus decoupling (1.3) and providing natural definitions for $\psi^\pm$ in regions where $n(\mathbf{x})$ (or S) is independent of x. We call this choice full *local* splitting, noting that it provides, in some sense, the most complete splitting. It is naturally used (and illustrated in this contribution) for media with deterministic $n(\mathbf{x})$ which varies with x. Clearly, as recognized previously,[2,4] there is no unique natural choice in regions where $n(\mathbf{x})$ varies with x. We now mention some other useful splitting choices. *Reference* splitting where $T(x) \equiv S_0$, independent of x, is also suitable for treating deterministic media where variations in $n(\mathbf{x})$ with respect to x are restricted to some localized region. Here we naturally choose $S_0 \equiv \lim_{|x| \to \infty} S(x)$. We have recently implemented reference splitting to treat wave propagation in random media where the (statistical) mean, $\langle n(\mathbf{x})\rangle$ of $n(\mathbf{x})$, is independent of x, and we choose $S_0 \equiv \langle S \rangle$.[7] Of course (1.2) and (1.3) also allow for the possibility of *intermediate* splittings where $T \not\equiv S$, but T still depends on x, e.g., $T(x) = \langle S(x) \rangle$ for random media where $\langle n(\mathbf{x})\rangle$ also varies with x.

Neglecting ± coupling in (1.3) produces a "unidirectional propagation approximation" which will be of the WKB (parabolic) type for local (reference) splitting. Such an approximation constitutes the lead term in an iterative Bremmer-type series expansion[1] of the exact solution of (1.3). For either an exact or approximate treatment, it is clearly necessary to develop an operational calculus for the splitting operator $T(x)$. This is trivial if one simply makes a scalar choice for $T(x)$ [e.g., $T(\mathbf{x}) \equiv k^2(x, \mathbf{x}_\perp = 0)$ (or the $|x| \to \infty$ limit, should it exist) which produces an Arnaud[3] (Leontovich–Fock[8]) approximation], but instead we con-

0022-2488/88/010097-06$02.50

sider only "more complete" abstract operator choices in an attempt to avoid "kinematic" contributions to backscattering [i.e., those not associated with variations of $k(\mathbf{x})$ with respect to x]. The spectral theory for $T(x)$ is only trivial for stratified media where T is multiplicative in the transverse Fourier transform variables. More generally, Weyl pseudodifferential operator calculus can be used,[9] but here we utilize conventional self-adjoint operator spectral theory, which, for focusing media or wave guides, corresponds to a wave field decomposition into a complete set of guided and radiation modes.

Mode-coupled equations, obtained by evaluating the abstract operator splitting equations (1.3) in a natural explicit representation, are displayed in Sec. II. A "more conventional" derivation of these equations is also provided. The explicit example of a square law medium with one lateral dimension is treated in Sec. III, and a diagrammatic representation of the Bremmer-type series solutions is provided. Invariant imbedding equations for transmission and reflection operators for slabs of finite thickness are presented in Sec. IV, and the symplectic structure of the underlying splitting equations is shown to generate important reciprocity conditions.

II. LOCAL SPLITTING APPLIED TO DETERMINISTIC FOCUSING MEDIA AND WAVEGUIDES

The infinite focusing media (or open waveguides) considered here have the following properties: (i) $n(\mathbf{x})$ attains its maximum near $\mathbf{x}_\perp = \mathbf{0}$; (ii) $n(\mathbf{x}) \to n_\infty(x)$, as $|\mathbf{x}_\perp| \to \infty$, for each x; and (iii) $n(\mathbf{x})$ is independent of x outside of the interval $(0<)\ a<x<b$. Thus any guided wave propagation is along the x axis, and scattering is restricted to $a<x<b$. The self-adjoint operator $T(x) \equiv S(x)$ here in general has several discrete eigenvalues satisfying $\lambda \in (k^2 n_\infty^2(x), k^2 \max n^2(\mathbf{x}))$. The corresponding $L^2(\mathbf{x}_\perp)$-normalized eigenfunctions describe the guided modes.[10,11] We note that if $d=3$ and $\delta n(\mathbf{x}) \equiv n(\mathbf{x}) - n_\infty(x) \in C_0^\infty(R^2)$ is non-negative, then there exists at least one such guided mode,[12] no matter how small k! (This is also a property of symmetric, but not asymmetric, slab waveguides[10].) In addition, each $\lambda \in [-\infty, k^2 n_\infty^2(x)]$ is in the continuous spectrum. Specifically, $\lambda_{\mathbf{k}_\perp} = k^2 n_\infty^2(x) - |\mathbf{k}_\perp|^2$ is associated with "weak" radiation mode eigenfunctions $\sim e^{i\mathbf{k}_\perp \cdot \mathbf{x}_\perp}$ as $|\mathbf{x}_\perp| \to \infty$. Modes with $\lambda > 0$ (<0) are described as propagating (evanescent) for reasons which will become obvious. A schematic of the spectrum of $T=S$ is shown in Fig. 1. Radiation modes can plan an important role in wave propagation, but one encounters fundamental problems associated with singularities in associated coupling terms[11] (see below). A guided mode can also disappear into the continuum of radiation modes as x varies, as a result of changes in the shape of $n(\mathbf{x})$. Such a *cutoff* highlights a fundamental problem with an "adiabatic" treatment neglecting mode coupling.[11] This problem will not be addressed here.

For a closed waveguide, $\mathbf{x}_\perp$ is restricted to a finite region for each x. Its boundary (where conditions are imposed on the wavefield) is assumed to vary smoothly with x for $a<x<b$, and to be fixed elsewhere. Here the spectrum of $T \equiv S$ is purely discrete, each eigenvalue λ corresponding to one or more guided modes. A mode which is propagating ($\lambda > 0$), for large x, could become evanescent for part of $a<x<b$ (the quantum mechanical analog of which is "barrier tunneling"[13]). In this case one sees singularities in the splitting procedure (certain $\lambda^{-1/2} \to \infty$) generating a strong coupling between forward and back propagating modes (cf. the connection formulas for barrier tunneling[13]). We shall not discuss this further here.

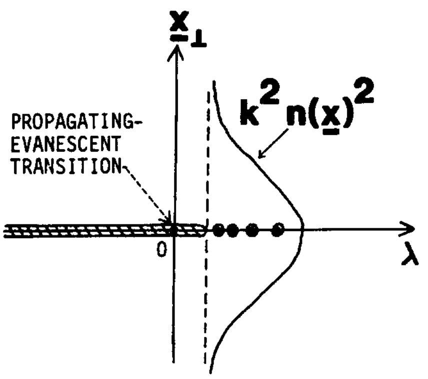

FIG. 1. A schematic of the continuum spectrum (cross-hatched line) and point spectrum (circles) of $T \equiv S$.

It is convenient to introduce generic mode labels κ and to denote all ($T \equiv S$)-mode eigenfunctions by $\psi_\kappa(\mathbf{x}_\perp | x) \equiv \langle \mathbf{x}_\perp | \kappa, x\rangle$ (using Dirac notation), and corresponding eigenvalues by $\lambda_\kappa(x)$. Thus if Σ_κ represents a sum/integral over all modes, then one has that

$$f(T(x)) = \sum_\kappa f(\lambda_\kappa(x)) |\kappa,x\rangle \langle \kappa,x| .$$

The modal coefficients $\phi_\kappa(x) = \langle \kappa,x|\phi\rangle$ of $\phi(\mathbf{x})$ satisfy

$$\phi(\mathbf{x}) \equiv \sum_\kappa \phi_\kappa(x) \langle \mathbf{x}_\perp | \kappa,x\rangle . \tag{2.1}$$

We note here that

$$(\psi_x)_\kappa = \left\langle \kappa,x \left| \frac{d}{dx} \right| \psi \right\rangle = \frac{d}{dx} \langle \kappa,x|\psi\rangle - \left(\frac{d}{dx} \langle \kappa,x| \right) |\psi\rangle$$
$$= \frac{d}{dx}\psi_\kappa + \sum_{\kappa'} \left\langle \kappa,x \left| \frac{d}{dx} \right| \kappa',x \right\rangle \psi_{\kappa'} , \tag{2.2}$$

so, from (1.2), one has

$$\psi_\kappa^\pm = \frac{1}{2}\left(\psi_\kappa \mp i\lambda_\kappa^{-1/2} \frac{d}{dx}\psi_\kappa \right)$$
$$\mp \frac{i}{2} \lambda_\kappa^{-1/2} \sum_{\kappa'} \left\langle \kappa,x \left| \frac{d}{dx} \right| \kappa',x \right\rangle \psi_{\kappa'} . \tag{2.3}$$

We can obtain *directly* from (1.3), with $T(x) \equiv S(x)$, coupled equations for $\psi_\kappa^\pm$, which after some rearrangement become

$$\frac{d}{dx}\psi_\kappa^\pm(x) + \left\{ \mp i\lambda_\kappa(x)^{1/2} + \frac{1}{4}\frac{d}{dx} \ln \lambda_\kappa(x) \right\} \psi_\kappa^\pm(x)$$
$$= \frac{1}{4}\frac{d}{dx} \ln \lambda_\kappa(x) \psi_\kappa^\mp(x)$$
$$+ \frac{1}{2} \sum_{\kappa'} \left\langle \kappa,x \left| \frac{d}{dx} \right| \kappa',x \right\rangle \left\{ \left[\left(\frac{\lambda_{\kappa'}(x)}{\lambda_\kappa(x)} \right)^{1/2} - 1 \right] \psi_{\kappa'}^\mp(x) \right.$$

$$- \left[\left(\frac{\lambda_{\kappa'}(x)}{\lambda_\kappa(x)} \right)^{1/2} + 1 \right] \psi_{\kappa'}^{\pm}(x) \Big\}$$

$$\equiv F_\kappa^{\pm}(x), \text{ say.} \tag{2.4}$$

Note that contributions to $(T(x)^{-1/2})_x$ come from the x dependence of both eigenvalues and eigenfunctions (here bras and kets). For evanescent $\lambda_\kappa < 0$, we set $\lambda_\kappa^{1/2} = i|\lambda_\kappa|^{1/2}$ guaranteeing that the corresponding components of ψ^+ (ψ^-) are exponentially decreasing as x increases (decreases). The singular behavior of the coupling coefficients, $\langle \kappa,x | d/dx | \kappa',x \rangle$, where κ, κ' are both radiation modes, is discussed in Appendix A for $d = 3$.

Clearly (2.4) provides a natural starting point for the analysis of backscattering effects on wave propagation. For boundary conditions corresponding to one or more right-propagating guided modes at $x = 0$ [$\psi_\kappa^+(0) \neq 0$, for such κ], and no left-propagating waves at $x = \infty$ [$\psi_\kappa^-(\infty) = 0$], (2.4) can be rewritten in integral form as

$$\psi_\kappa^+(x) = \phi_\kappa^+(x) + \int_0^x dx'\, G_\kappa^+(x|x') F_\kappa^+(x'),$$
$$\psi_\kappa^-(x) = -\int_x^\infty dx'\, G_\kappa^-(x|x') F_\kappa^-(x'), \tag{2.5}$$

where

$$\phi_\kappa^+(x) \equiv G_\kappa^+(x|0) \psi_\kappa^+(0)$$

and

$$G_\kappa^{\pm}(x|x') \equiv \left(\frac{\lambda_\kappa(x')}{\lambda_\kappa(x)} \right)^{1/4} \exp\left(\pm i \int_{x'}^x dx''\, \lambda_\kappa^{1/2}(x'') \right).$$

The only contribution to the integrals, associated with inhomogeneity in $n(\mathbf{x})$ with respect to x, comes from the scattering region $x' \in [a,b]$. If coupling between guided modes is weak and coupling to radiation modes can be ignored, then the iterative solution of (2.5) is viable.

It is instructive to consider the relationship of (2.4) to the more conventional mode-coupled equations for ψ_κ, $(\psi_x)_\kappa$. We show, in Appendix B, how the latter can be used to generate a standard second-order equation for the ψ_κ [as could have been obtained from an explicit propagation mode representation of (1.1)]. By introducing an appropriate infinite matrix splitting operator, we can also recover (2.4).

III. WAVE PROPAGATION IN SQUARE LAW MEDIA (WITH VARIABLE FOCUSING)

When the guided mode wave propagation in focusing media is effectively confined laterally to a region near the maximum of $n(\mathbf{x})$, one might expect a quadratic approximation for $n(\mathbf{x})$ to be reasonable. This motivates the analysis of "square-law" media where

$$n(\mathbf{x})^2 = 1 - B^2(x)|\mathbf{x}_\perp|^2, \tag{3.1}$$

which, of course, is unphysical for $|\mathbf{x}_\perp| > B^{-1}$. Relation (3.1) provides a useful description for certain optical fibers. Although replacing the physical $n(\mathbf{x})$ by (3.1) may have minimal effect on the highest (guided mode) eigenvalues and eigenfunctions of $T(x) \equiv S(x)$ and the corresponding eigenfunctions, it affects those of lower eigenvalues more dramatically, and replaces the continuous radiation mode spectrum with a "spurious" point spectrum.

For simplicity we confine our attention to $d = 2$ here (a single lateral dimension). Here the eigenfunctions and eigenvalues of

$$T(x) \equiv S(x) = \frac{\partial^2}{\partial x_\perp{}^2} + k^2(1 - B^2(x) x_\perp{}^2)$$

are given by

$$\psi_m(x_\perp | x) = (2^{-m}/m!)^{1/2} (kB(x)/\pi)^{1/4}$$
$$\times H_m(k^{1/2} B(x)^{1/2} x_\perp) e^{-kB(x)x_\perp^2/2}, \tag{3.2}$$
$$\lambda_m(x) = k^2 - 2kB(x)(m + \tfrac{1}{2}), \quad \text{for } m \geqslant 0,$$

where H_m is the mth-order Hermite polynomial. Using standard relationships for the H_m, one can show that

$$\frac{d}{dx} |m,x\rangle = \frac{B'(x)}{4B(x)} [m^{1/2}(m-1)^{1/2} |m-2,x\rangle$$
$$- (m+2)^{1/2}(m+1)^{1/2} |m+2,x\rangle], \tag{3.3a}$$

so

$$\left\langle m,x \left| \frac{d}{dx} \right| n,x \right\rangle = \frac{B'(x)}{4B(x)} [(m+2)^{1/2}(m+1)^{1/2} \delta_{m+2,n}$$
$$- m^{1/2}(m-1)^{1/2} \delta_{m-2,n}]. \tag{3.3b}$$

It is elucidating to consider the high wavenumber (k) regime here where $d/dx \ln \lambda_m$ and $(\lambda_{m\pm 2}/\lambda_m)^{1/2} - 1 = O(1/k)$, which indicates the small coupling between forward and backward propagating modes. In this regime (2.4) becomes

$$\frac{d}{dx} \psi_m^{\pm} \mp i \left[k - B\left(m + \frac{1}{2} \right) \right] \psi_m^{\pm}$$
$$= \frac{B'(x)}{4B(x)} [m^{1/2}(m-1)^{1/2} \psi_{m-2}^{\pm}$$
$$- (m+2)^{1/2}(m+1)^{1/2} \psi_{m+2}^{\pm}] + O\left(\frac{1}{k} \right). \tag{3.4}$$

Let us now utilize the integral form (2.5) of the mode coupled equations (2.4) for a scattering problem with boundary conditions $\psi_m^+(0) \propto \delta_{m,0}$, $\psi_m^-(\infty) = 0$ for all $m \geqslant 0$. Clearly, from (3.3) and (2.4), one has that $\psi_m^{\pm}(x) \equiv 0$, for m odd. Expressions for $\psi_m^{\pm}$, with m even, can be obtained from the iterative solution of (2.5) [assuming that no $\lambda_m(x)$ changes sign or becomes zero, as x varies]. It is natural to represent contributions to these solutions diagrammatically in terms of paths on a lattice of points labeled by the modes $(m, \pm)$. The zero length path $(0, +)$ and segments connecting different points have the interpretation shown in Fig. 2. Then $\psi_m^{\pm}$ is represented as a sum over all paths connecting $(0, +)$ to $(m, \pm)$ (see Fig. 3). One can straightforwardly extend these considerations to higher dimensional ($d \geqslant 3$) square law media.

IV. INVARIANT IMBEDDING, SYMPLECTIC STRUCTURE AND RECIPROCITY, AND OTHER SYMMETRIES

We have shown that the basic differential equation associated with any splitting has the form

$$\overset{(0,+)}{\bullet} = (\lambda_0(0)/\lambda_0(x))^{1/4}\exp[i\int_0^x dx'\ \lambda_0(x')^{1/2}]\psi_0^+(0)$$

$$\overset{(m,+)}{\circ}\!\longrightarrow\!\overset{(m\pm2,\sigma)}{\bullet} = \mp\frac{\sigma}{8}(m+1\pm1)^{1/2}(m\pm1)^{1/2}\int_0^x dx' G_m^+(x|x')\left[\left(\frac{\lambda_{m\mp2}(x')}{\lambda_m(x')}\right)^{1/2}+\sigma\right]\frac{B'(x')}{B(x')}\ \bullet$$

$$\overset{(m,+)}{\circ}\!\longrightarrow\!\overset{(m,-)}{\bullet} = -\frac{2m+1}{4}k\int_0^x dx' G_m^+(x|x')B'(x')/\lambda_m(x')\bullet$$

$\overset{(m,-)}{\circ}\!\longrightarrow\!\overset{(m\pm2,-\sigma)}{\bullet}$ $\left(\overset{(m,-)}{\bullet}\!\longrightarrow\!\overset{(m,+)}{\bullet}\right)$ are obtained from $\overset{(m,+)}{\circ}\!\longrightarrow\!\overset{(m\pm2,\sigma)}{\bullet}$ $\left(\overset{(m,+)}{\circ}\!\longrightarrow\!\overset{(m,-)}{\bullet}\right)$

by replacing $\int_0^x dx'\ G_m^+\ldots$ with $\int_x^\infty dx'\ G_m^-\ldots$

FIG. 2. Operator theoretic interpretation of path segments appearing in the diagrammatic representation of solutions of the coupled wave equations. Here $\sigma = +1$ or -1.

$$\frac{d}{dx}\begin{pmatrix}\psi^+\\ \psi^-\end{pmatrix} = \mathsf{A}(x)\begin{pmatrix}\psi^+\\ \psi^-\end{pmatrix} \equiv \mathsf{j}\mathsf{H}(x)\begin{pmatrix}\psi^+\\ \psi^-\end{pmatrix},$$

$$\text{where } \mathsf{j} = \begin{pmatrix}0 & I\\ -I & 0\end{pmatrix}, \tag{4.1}$$

defining $\mathsf{H}(x) \equiv -\mathsf{j}\mathsf{A}(x)$ and noting that $-\mathsf{j}^2 = \mathsf{I}$ (the identity). Since (4.1) is linear, one naturally defines the abstract transmission $T^\pm$ and reflection $R^\pm$ operators for slabs $[x,y]$ of finite thickness, by

$$\begin{pmatrix}\psi^-(x)\\ \psi^+(y)\end{pmatrix} = \begin{pmatrix}R^+(x,y) & T^-(x,y)\\ T^+(x,y) & R^-(x,y)\end{pmatrix}\begin{pmatrix}\psi^+(x)\\ \psi^-(y)\end{pmatrix}$$

$$\equiv \mathsf{S}(x,y)\begin{pmatrix}\psi^+(x)\\ \psi^-(y)\end{pmatrix}, \tag{4.2}$$

where S is called the scattering operator and clearly $T^\pm(x,x) = I$, $R^\pm(x,x) = 0$. The operator S satisfies the differential equation (cf. Refs. 4, 5, and 14)

$$\frac{\partial}{\partial y}\mathsf{S} = \begin{pmatrix}T^- & 0\\ R^- & I\end{pmatrix}\mathsf{H}(y)\begin{pmatrix}T^+ & R^-\\ 0 & I\end{pmatrix}. \tag{4.3}$$

Taking the four components of (4.3) provides the familiar Ambarzumian form of the invariant imbedding equations.[14] An equivalent set may be obtained from these by making the replacements $\partial/\partial y \to \partial/\partial x$, $T^- \leftrightarrow T^+$, $R^- \leftrightarrow R^+$, $H_{\pm\pm}(y) \to H_{\mp\mp}(x)$, $H_{\pm\mp}(y) \to H_{\mp\pm}(x)$.

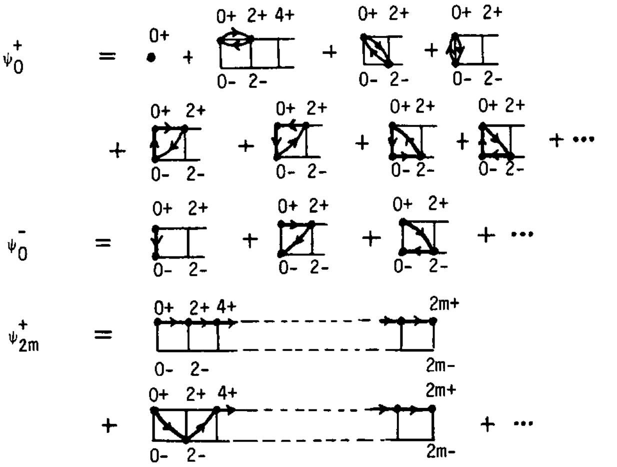

FIG. 3. Diagrammatic expansions for various forward and backward traveling modal components of the wave field.

Since these equations have a structure generic to many problems in wave propagation and transport theory, we anticipate that there exist basic relationships between the reflection and transmission operators. To fully elucidate this structure, it is appropriate to introduce several new quantities. Let $\mathsf{C}(x,y)$ be the operator which propagates the wavefields $\psi^\pm$ from x to y, i.e.,

$$\begin{pmatrix}\psi^+(y)\\ \psi^-(y)\end{pmatrix} = \mathsf{C}(x,y)\begin{pmatrix}\psi^+(x)\\ \psi^-(x)\end{pmatrix}, \tag{4.4}$$

where, from (4.2),

$$C_{++} = T^+ - R^-[T^-]^{-1}R^+,\quad C_{+-} = R^-[T^-]^{-1},$$

$$C_{-+} = -[T^-]^{-1}R^+,\quad C_{--} = [T^-]^{-1}. \tag{4.5}$$

Though C is less physical than S, we shall see that in certain cases it can be regarded as a (linear) canonical transformation. Note that from (4.1) and (4.4), one clearly has

$$\mathsf{C}(x,x+\Delta x) = \mathsf{I} + \mathsf{A}(x)\Delta x + O(\Delta x^2), \tag{4.6}$$

where I is the identity. Finally, it is convenient to define

$$\boldsymbol{\theta}^1(x) = \begin{pmatrix}0 & \theta(x)\\ -\tilde\theta(x) & 0\end{pmatrix},\quad \boldsymbol{\theta}(x,y) = \begin{pmatrix}\theta(x) & 0\\ 0 & \tilde\theta(y)\end{pmatrix}, \tag{4.7}$$

where the operator $\theta(x)$ will be specified later, and $\tilde{}$ denotes a *real* involution operation (so $\tilde{\tilde A} = A$, $\tilde i = i$). Now using (4.1)–(4.7) as defining relations, one has the following.

Theorem: The following conditions are equivalent for any differentiable $\theta(x)$:

$$\text{(i) } \boldsymbol{\theta}(x,y)\mathsf{S}(x,y) = \tilde{\mathsf{S}}(x,y)\tilde{\boldsymbol{\theta}}(x,y), \tag{4.8}$$

i.e.,

$$\theta(x)R^+(x,y) = \tilde R^+(x,y)\tilde\theta(x),$$

$$\tilde\theta(y)R^-(x,y) = \tilde R^-(x,y)\theta(y)$$

and

$$\theta(x)T^-(x,y) = \tilde T^+(x,y)\theta(y);$$

$$\text{(ii) } \tilde{\mathsf{C}}(x,y)\boldsymbol{\theta}^1(y)\mathsf{C}(x,y) = \boldsymbol{\theta}^1(x), \tag{4.9}$$

i.e., a symplectic condition for C;

$$\text{(iii) } \tilde{\mathsf{A}}(x)\boldsymbol{\theta}^1(x) + \boldsymbol{\theta}^1(x)\mathsf{A}(x) + \boldsymbol{\theta}_x^1(x) = \mathsf{O}, \tag{4.10}$$

or equivalently,

$$\tilde{\mathsf{H}}(x)\tilde{\boldsymbol{\theta}}(x,x) - \boldsymbol{\theta}(x,x)\mathsf{H}(x) + \boldsymbol{\theta}_x^1(x) = 0,$$

i.e.,

$$\tilde H_{++}(x)\tilde\theta(x) = \theta(x)H_{++}(x),$$

$$\tilde H_{--}(x)\theta(x) = \tilde\theta(x)H_{--}(x),$$

and

$$\tilde H_{-+}(x)\theta(x) - \theta(x)H_{+-}(x) + \theta_x(x) = 0.$$

Proof: (i) $\Rightarrow$ (ii): Calculation of the components of $\tilde{\mathsf{C}}(x,y)\boldsymbol{\theta}^1(y)\mathsf{C}(x,y)$, followed by substitution of identities from (i), shows straightforwardly that this quantity equals $\boldsymbol{\theta}^1(x)$.

(ii) $\Rightarrow$ (iii): Substituting the expansions

$$\tilde{\mathsf{C}}(x,x+\Delta x) = \mathsf{I} + \tilde{\mathsf{A}}\Delta x + O(\Delta x^2),$$

$$\mathsf{C}(x,x+\Delta x)^{-1} = \mathsf{I} - \mathsf{A}\Delta x + O(\Delta x^2), \tag{4.11}$$

and

$$\boldsymbol{\theta}^1(x+\Delta x)=\boldsymbol{\theta}^1(x)+\boldsymbol{\theta}^1_x(x)\Delta x+O(\Delta x^2)\,,$$

into the identity

$$\widetilde{\mathsf{C}}(x,x+\Delta x)\boldsymbol{\theta}^1(x+\Delta x)=\boldsymbol{\theta}^1(x)\mathsf{C}(x,x+\Delta x)^{-1}\,, \tag{4.12}$$

and equating terms $O(\Delta x)$ yields (iii).

(iii) ⇒ (i): Using Eqs. (4.3) and identities from (iii), one obtains

$$\begin{aligned}\frac{\partial}{\partial y}[\theta(x)R^+]&=\theta(x)T^-H_{++}(y)T^+\\&=\widetilde{T}^+\theta(y)H_{++}(y)T^+\\&\quad+[\theta(x)T^--\widetilde{T}\theta(y)]H_{++}(y)T^+\,,\end{aligned}$$

and

$$\begin{aligned}\frac{\partial}{\partial y}[\widetilde{R}^+\tilde{\theta}(x)]&=\widetilde{T}^+\widetilde{H}_{++}(y)\widetilde{T}^-\tilde{\theta}(x)\\&=\widetilde{T}^+\widetilde{H}_{++}(y)\tilde{\theta}(y)T^++\widetilde{T}^+\widetilde{H}_{++}(y)\\&\quad\times[\widetilde{T}^-\tilde{\theta}(x)-\tilde{\theta}(y)T^+]\,,\end{aligned} \tag{4.13}$$

so

$$\begin{aligned}&\frac{\partial}{\partial y}[\theta(x)R^+-\widetilde{R}^+\tilde{\theta}(x)]\\&\quad=[\theta(x)T^--\widetilde{T}^+\theta(y)]H_{++}(y)T^+-I\,,\end{aligned} \tag{4.14}$$

where I denotes the involution of the first term. Similarly,

$$\begin{aligned}&\frac{\partial}{\partial y}[\tilde{\theta}(y)R^--\widetilde{R}^-\theta(y)]\\&\quad=[\tilde{\theta}(y)R^--\widetilde{R}^-\theta(y)]\\&\quad\quad\times[H_{++}(y)\overline{R}+H_{++}(y)]-I\,,\end{aligned} \tag{4.15}$$

$$\begin{aligned}&\frac{\partial}{\partial y}[\tilde{\theta}(y)T^+-\widetilde{T}^-\tilde{\theta}(x)]\\&\quad=[\tilde{\theta}(y)R^--\widetilde{R}^-\theta(y)]H_{++}(y)T^+\\&\quad\quad-[\widetilde{R}^-\widetilde{H}_{++}(y)+\widetilde{H}_{++}(y)]\\&\quad\quad\times[\widetilde{T}^-\tilde{\theta}(x)-\tilde{\theta}(y)T^+]\,.\end{aligned} \tag{4.16}$$

Since the identities (i) [i.e., (4.8)] are trivially satisfied when $x=y$, (4.14)–(4.16) show that they are satisfied for all $y\geqslant x$. □

Now we apply these results to the specific choice of splitting of ψ into $\psi^\pm$ defined by (1.2) and thus associated with the operator $T=T(x)$. The corresponding components of H can be determined from (1.3). For this application it is necessary to choose the real involution $\tilde{}$ to correspond to the real transpose (rather than Hermitian adjoint) and to note that appropriate choices of T satisfy $\widetilde{T}=T$, i.e.,

$$\begin{aligned}\int d\mathbf{x}_\perp\,\psi(\mathbf{x}_\perp)(\widetilde{T}\phi)(\mathbf{x}_\perp)&\equiv\int d\mathbf{x}_\perp\,\phi(\mathbf{x}_\perp)(T\psi)(\mathbf{x}_\perp)\\&=\int d\mathbf{x}_\perp\,\psi(\mathbf{x}_\perp)(T\phi)(\mathbf{x}_\perp)\,.\end{aligned} \tag{4.17}$$

This is obviously true choosing, e.g., $T=S(x)=\Delta_\perp+k^2(\mathbf{x})$ (local splitting) or $T=S_0=\Delta_\perp+k^2(x=\pm\infty,\mathbf{x}_\perp)$ (reference splitting) *even if* $k(\mathbf{x})=kn(\mathbf{x})$ is complex valued corresponding to a dissipative medium. Then condition (iii) is satisfied by the choice

$$\theta=T^{1/2}\,, \tag{4.18}$$

as may be verified by straightforward calculation.

Another symmetry property for the operator $\mathsf{C}(x,y)$ is based on the observation that the (easily verified) relationship

$$\begin{pmatrix}0&I\\I&0\end{pmatrix}\overline{\mathsf{A}}(x)=\mathsf{A}(x)\begin{pmatrix}0&I\\I&0\end{pmatrix},$$

or

$$A_{++}=\overline{A}_{--},\quad A_{-+}=\overline{A}_{+-}\,, \tag{4.19}$$

is equivalent to

$$\begin{pmatrix}0&I\\I&0\end{pmatrix}\overline{\mathsf{C}}(x,y=\mathsf{C}(x,y))\begin{pmatrix}0&I\\I&0\end{pmatrix},$$

or

$$C_{++}(x,y)=\overline{C}_{--}(x,y),\quad C_{-+}(x,y)=\overline{C}_{+-}(x,y)\,. \tag{4.20}$$

This is representative of a broader class of symmetry relationships.[15] When (4.20) is combined with (4.9) and (4.18), one also obtains

$$\begin{aligned}&\widetilde{\mathsf{C}}(x,y)\begin{pmatrix}T^{1/2}(y)&0\\0&-T^{/2}(y)\end{pmatrix}\overline{\mathsf{C}}(x,y)\\&\quad=\begin{pmatrix}T^{1/2}(x)&0\\0&-T^{1/2}(x)\end{pmatrix}.\end{aligned} \tag{4.21}$$

Matrix elements of scattering operators are evaluated here using a natural mixed representation with respect to eigenfunctions of (different positioned) splitting operators $T(x)$. For example, $T^+_{\kappa',\kappa}(x,y)=\langle\kappa',y|T^+(x,y)|\kappa,x\rangle$ is the appropriate transmission coefficient connecting right propagating modes κ at x, and κ' at y. Generic Dirac notation is used here for T eigenbras and eigenkets, and corresponding eigenvalues are denoted by $\lambda_\kappa(x)$ (but now these will *not* correspond to S eigenbras and eigenkets and eigenvalues when $S\neq T$). This prescription is automatically compatible with the evaluation of operator products required in (4.3) (or equivalent versions of these equations). Clearly, in (4.3), T bras and kets for all components of $\mathsf{H}(y)$ are evaluated at y. The important reciprocity conditions (4.8) [using (4.18)] have the explicit form

$$\begin{aligned}&\lambda_\kappa(x)^{1/2}\langle\kappa,x|R^+(x,y)|\kappa',x\rangle\\&\quad=\lambda_{\kappa'}(x)^{1/2}\langle\bar{\kappa}',x|R^+(x,y)|\bar{\kappa},x\rangle\,,\\&\lambda_\kappa(y)^{1/2}\langle\kappa,y|R^-(x,y)|\kappa',y)\\&\quad=\lambda_{\kappa'}(y\rangle^{1/2}\langle\bar{\kappa}',y|R^+(x,y)|\bar{\kappa},y\rangle\,,\\&\lambda_\kappa(x)^{1/2}\langle\kappa,x|T^-(x,y)|\kappa',y\rangle\\&\quad=\lambda_{\kappa'}(y)^{1/2}\langle\bar{\kappa}',y|T^+(x,y)|\bar{\kappa},x\rangle\,,\end{aligned} \tag{4.22}$$

where $\langle\mathbf{x}_\perp|\bar{\kappa},x\rangle=\psi_\kappa(\mathbf{x}_\perp)^*$.

It is a straightforward matter to write down the explicit form of the mode coupled invariant imbedding equations. One could investigate an iterative form of solution which, to the lowest order, gives

$$T^\pm_{\kappa',\kappa}\sim\delta_{\kappa',\kappa}\exp\left(\int_x^y ds\langle\kappa,s|H_{\mp\pm}|\kappa,s\rangle\right)\text{ and }R^\pm_{\kappa',\kappa}\sim 0\,. \tag{4.23}$$

V. CONCLUSIONS

An abstract splitting operator based formation is shown to provide a powerful and flexible formulation of wave propagation in "imperfect" media. Mode coupled equations connecting forward and backward propagation provide a natural basis for the analysis of backscattering effects. We have, however, noted some difficulties associated with guided mode cutoff, and propagating-evanescent transitions. The formalism also provides a natural basis for derivation of invariant imbedding equations for transmission and reflection operators. The reciprocity relations derived here for these are important from a fundamental and practical perspective.

ACKNOWLEDGMENTS

The author would like to thank Professor J. P. Corones for making available his unpublished notes on reciprocity and symmetries for invariant imbedding equations, and for his encouragement and interest in this work.

This work was supported by the Applied Mathematical Sciences subprogram of the Office of Energy Research, U.S. Department of Energy, under Contract No. W-7405-ENG-82.

APPENDIX A: RADIATION MODE EIGENFUNCTIONS AND MATRIX ELEMENTS FOR $d=3$

Here the eigenvalue equation for the radiation mode eigenfunctions, $\psi_{\mathbf{k}_\perp} \sim (1/2\pi)e^{i\mathbf{k}_\perp\cdot\mathbf{x}_\perp}$, as $|\mathbf{x}_\perp| \to \infty$, can be converted to the integral form

$$\psi_{\mathbf{k}_\perp}(\mathbf{x}_\perp|x) = \frac{1}{2\pi}e^{i\mathbf{k}_\perp\cdot\mathbf{x}_\perp} + \frac{ik^2}{4}\int d\mathbf{x}'_\perp \times H^1_0(k_\perp|\mathbf{x}_\perp - \mathbf{x}'_\perp|)\delta n(x,\mathbf{x}'_\perp)\psi_{\mathbf{k}_\perp}(\mathbf{x}'_\perp|x)\,, \tag{A1}$$

where $k_\perp = |\mathbf{k}_\perp|$, and we have used the Hankel function H^1_0 to provide an explicit representation of the two-dimensional free Green's function $(\Delta_\perp + k^2_\perp)^{-1}$ (see Ref. 16a). Let us analyze radiation to radiation mode coupling coefficients, $\langle \mathbf{k}_\perp,x|d/dx|\mathbf{k}'_\perp,x\rangle$, of (2.4). First, one must consider $d/dx\,\psi_{\mathbf{k}'_\perp}$, which can be obtained from (A1) by differentiating under the integral sign. Thus its large $x_\perp = |\mathbf{x}_\perp|$ asymptotic behavior is obtained directly from that of

$$H^1_0(k'_\perp x_\perp) \sim (\pi k'_\perp x_\perp/2)^{-1/2}\exp(ik'_\perp x_\perp - i\pi/4)\,.$$

Second, it is convenient to reexpress the plane wave part of $\psi_{\mathbf{k}_\perp}$ as a linear combination of cylindrical wave eigenfunctions of $\Delta_\perp$, proportional to

$$J_\nu(k_\perp x_\perp) \sim \left(\frac{\pi k_\perp x_\perp}{2}\right)^{-1/2}\cos\left(k_\perp x_\perp - \frac{1}{2\nu\pi} - \frac{1}{4\pi}\right),$$

as $x_\perp \to \infty$.[16b] After writing

$$\int d\mathbf{x}_\perp = \int d\phi \int dx_\perp\, x_\perp\cdot,$$

it is clear that these coupling coefficients involve singular integrals of the form

$$\int_0^\infty dk\, e^{ikx} = \frac{1}{2}\delta(k) + i\mathbb{P}/k\,, \tag{A2}$$

where $\mathbb{P}$ represents a Cauchy principal value integral.

APPENDIX B: SPLITTING OF CONVENTIONAL MODE-COUPLED EQUATIONS

Let Ψ, Ψ_x denote infinite dimensional vectors with components ψ_κ, $(\psi_x)_\kappa$, respectively. Then, from (1.1), one can readily obtain the following infinite matrix form of the conventional mode-coupled equations[11]:

$$\frac{d}{dx}\begin{pmatrix}\Psi\\ \Psi_x\end{pmatrix} = \left(\begin{array}{c:c} -\mathsf{D} & \mathsf{I} \\ \hdashline -\lambda & -\mathsf{D}\end{array}\right)\begin{pmatrix}\Psi\\ \Psi_x\end{pmatrix}, \tag{B1}$$

where $(\mathsf{I})_{\kappa,\kappa'} = \delta_{\kappa,\kappa'}$ is the identity, $(\lambda)_{\kappa,\kappa'} = \delta_{\kappa,\kappa'}\lambda_\kappa$ (where the λ_κ are the eigenvalues of $T \equiv S$), and $(\mathsf{D})_{\kappa,\kappa'} = \langle\kappa,x|d/dx|\kappa',x\rangle$. Elimination of Ψ_x from (B1) yields the standard second-order equation for Ψ[11]:

$$\frac{d^2}{dx^2}\Psi + 2\mathsf{D}\frac{d}{dx}\Psi + \left(\mathsf{D}^2 + \frac{d}{dx}\mathsf{D} + \lambda\right)\Psi = 0\,. \tag{B2}$$

Instead we introduce right, Ψ^+, and left, Ψ^-, traveling vectors in terms of a splitting operator P by

$$\begin{pmatrix}\Psi^+\\ \Psi^-\end{pmatrix} = \mathsf{P}\begin{pmatrix}\Psi\\ \Psi_x\end{pmatrix}, \quad \text{where } \mathsf{P} = \frac{1}{2}\left(\begin{array}{c:c} \mathsf{I} & -i\lambda^{1/2} \\ \hdashline \mathsf{I} & +i\lambda^{1/2}\end{array}\right), \tag{B3}$$

so the components of $\Psi^\pm$ are just $\psi^\pm_\kappa$ (for local splitting where $T \equiv S$). Deriving equations for $\Psi^\pm$ from (B1) in the obvious way [cf. (1.3)] yields

$$\begin{aligned}\frac{d}{dx}\Psi^\pm = {} & \pm i\lambda^{1/2}\Psi^\pm + \frac{1}{2}(\lambda^{-1/2})_x\lambda^{1/2}(\Psi^\pm - \Psi^\mp)\,,\\ & -\tfrac{1}{2}\lambda^{-1/2}\mathsf{D}\lambda^{1/2}(\Psi^\pm - \Psi^\mp)\\ & -\tfrac{1}{2}\mathsf{D}(\Psi^+ + \Psi^-)\,,\end{aligned} \tag{B4}$$

recovering (2.4).

[1] H. Bremmer, Commun. Pure Appl. Math. **4**, 105 (1951).

[2] F. W. Sluijter, J. Math. Anal. Appl. **27**, 282 (1969); J. Opt. Soc. Am. **60**, 8 (1970).

[3] J. A. Arnaud, Bell Syst. Tech. J. **49**, 2311 (1970).

[4] J. Corones, J. Math. Anal. Appl. **50**, 361 (1975).

[5] M. E. Davison, "A general approach to splitting and invariant imbedding techniques for linear wave equations," Ames Laboratory preprint, 1982.

[6] J. P. Corones and R. J. Krueger, J. Math. Phys. **24**, 2301 (1983).

[7] J. W. Evans, J. Math. Phys. **26**, 2196 (1985).

[8] V. A. Fock, *Electromagnetic Diffraction and Propagation Problems* (McMillan, New York, 1960).

[9] L. Fishman and J. J. McCoy, Proc. Soc. Photo Opt. Instrum. Eng. **358**, 168 (1982); J. Math. Phys. **25**, 285 (1984).

[10] D. Marcuse, *Light Transmission Optics* (Van Nostrand, New York, 1972); *Theory of Optical Dielectric Waveguides* (Academic, New York, 1974).

[11] J. M. Arnold, in *Hybrid Formulation of Wave Propagation and Scattering*, edited by L. B. Felsen (Nijhoff, Dordrecht, 1984).

[12] M. Reed and B. Simon, *Methods of Modern Mathematical Physics: IV Analysis of Operators* (Academic, New York, 1978), p. 100.

[13] E. Merzbacher, *Quantum Mechanics* (Wiley, New York, 1961).

[14] G. M. Wing, *An Introduction to Transport Theory* (Wiley, New York, 1962); R. Bellman and R. Vasudevan, *Wave Propagation: An Invariant Imbedding Approach* (Reidel, Dordrecht, 1986).

[15] For any differentiable $\Phi(x)$, the following are equivalent:

(i) $\Phi(y)\overline{\mathsf{C}}(x,y) = \mathsf{C}(x,y)\Phi(x)$,

(ii) $\Phi(x)\overline{\mathsf{A}}(x) - \mathsf{A}(x)\Phi(x) + \Phi_x(x) = 0$.

[16] P. M. Morse and H. Feshbach, *Methods of Theoretical Physics, Part 1* (McGraw-Hill, New York, 1953), (a) pp. 810 and 822; (b) p. 828.

The *Zitterbewegung* of a Dirac particle in two-dimensional space-time

Takashi Ichinose and Hiroshi Tamura
Department of Mathematics, Kanazawa University, 920-Kanazawa, Japan

(Received 28 April 1987; accepted for publication 16 September 1987)

The path space measures that have been constructed in the present authors' previous papers to give path integral formulas in quantum mechanics for a Dirac particle in two-dimensional space-time are shown to be concentrated on those paths that have differential coefficients equal to plus or minus the light velocity in every finite time interval except at finitely many instants of time.

I. INTRODUCTION

In previous papers,[1–3] we presented a mathematically rigorous treatment of the Feynman path integral in relativistic quantum mechanics in two space-time dimensions. We constructed countably additive path space measures to give path integral formulas that represent the fundamental solution of the Cauchy problem for the Dirac equation as well as the retarded and the advanced propagators for a Dirac particle, both in the presence of an electromagnetic field. It was shown that the path space measures obtained have support on the sets of those Lipschitz continuous paths whose differential coefficients are of magnitude smaller than or equal to the light velocity at almost every instant of time (on the sets of the straight lines with slopes equal to plus or minus the light velocity, when the mass of the particle is zero).

In Ref. 4, we have improved the above result on their support property to show that almost every path has a differential coefficient that is, in magnitude, constant and exactly equal to the light velocity with the possible exception of a closed subset of time of Lebesgue measure zero.

The aim of the present paper is to give the ultimate result on the support property of these path space measures, that is, that they are concentrated on the sets of those Lipschitz continuous paths which have differential coefficients of magnitude equal to the light velocity in every finite time interval except at most finitely many instants of time. So the trajectory of the particle shuttles back and forth in one-dimensional space with slopes of the light velocity; it is a zigzag path of a finite number of straight segments in each finite time interval. At the end points of the segments, the particle changes its direction of motion. This property is related to the cryptic description of Feynman–Hibbs[5] (see also Riazanov[6] and Rosen[7]), and may be considered as a measure-theoretic interpretation of the notion of *Zitterbewegung*[8] of a Dirac particle in two-dimensional space-time.

The recent work by Blanchard *et al.*[9] has dealt with path integral formulas for the Diract equation based on the Poisson process.[10,11] In their approach, the support property corresponding to the result of the present paper is a direct consequence of the properties of the Poisson process. Our result is, however, a direct and analytic derivation from the path space measures constructed, not passing through the Poisson process.

In Sec. II, we first give a brief review of the result in our previous papers,[3,4] and then state the result of this paper on the support property of the path space measures constructed. Section III is devoted to its proof.

Here $\mathbb{C}^2$ is the vector space of complex two-column vectors, $(\mathbb{C}^2)'$ that of complex two-row vectors, and $M_2(\mathbb{C})$ that of complex 2×2 matrices.

II. RESULT

Consider two hyperbolic systems of the first order. One is the Dirac equation

$$\partial_t\phi(t,x) = -[\alpha(\partial_x - ieA(t,x)) + im\beta + ie\Phi(t,x)]\phi(t,x), \quad t\in\mathbb{R}, \quad x\in\mathbb{R}, \tag{2.1}$$

for a particle of mass m and charge e in an external electromagnetic field in two-dimensional space-time $\mathbb{R}\times\mathbb{R} = \mathbb{R}^2$. Here the real-valued functions $\Phi(t,x)$ and $A(t,x)$ are the scalar and vector potentials of the field, and α and β are 2×2 Hermitian matrices with $\alpha^2 = \beta^2 = 1$ and $\alpha\beta + \beta\alpha = 0$. The other system is

$$\begin{aligned}\partial_\tau\psi(\tau,\mathbf{x}) &= -i(H\psi)(\tau,\mathbf{x})\\ &\equiv -[\partial_0 + ieA_0(\mathbf{x}) + \alpha(\partial_1 + ieA_1(\mathbf{x}))\\ &\quad + im\beta]\psi(\tau,\mathbf{x}),\\ &\quad \tau\in\mathbb{R}, \quad \mathbf{x} = (x^0,x^1)\in\mathbb{R}^2,\end{aligned} \tag{2.2}$$

where $\partial_\rho = \partial/\partial x^\rho$, $\rho = 0,1$, and $A_0(\mathbf{x}) = \Phi(x^0,x^1)$, $A_1(\mathbf{x}) = -A(x^0,x^1)$ with $\mathbf{x} = (x^0,x^1)$ replacing (t,x). Here τ is a third variable, which may be regarded as a fictitious time. Equation (2.2) was used to construct the path space measure for path integral representations of the retarded and advanced propagators of the Dirac particle.[2,3] In these equations the natural units are used in which the light velocity c and the reduced Planck's constant $\hbar$ are equal to 1.

Then for the Cauchy problems for Eqs. (2.1) and (2.2) with data $\phi(r,x) = g(x)$ and $\psi(r,\mathbf{x}) = g(\mathbf{x})$, we have established[3] the following path integral formulas representing their respective solutions $\phi(t,x)$ and $\psi(t,\mathbf{x})$ as well as fundamental solutions $K^{\mathrm{I}}(s,x;r,y)$ and $K^{\mathrm{II}}(s,\mathbf{x};r,\mathbf{y})$. By $|r,s|$ we denote the closed interval $[r,s]$ when $r<s$, or $[s,r]$ when $r>s$.

Path integral representation: (1) There exists a unique $\mathscr{S}'(\mathbb{R}\times\mathbf{R};M_2(\mathbb{C}))$-valued countably additive measure $\nu^{\mathrm{I}}_{s;r}$ on the Banach space $C(|r,s|;\mathbf{R})$ of the one-dimensional continuous paths X: $|r,s|\to\mathbb{R}$ such that for every continuous $A(t,x)$ and $\Phi(t,x)$,

0022-2488/88/010103-07$02.50

$$(f,\phi(s,\cdot)) = \int\int_{\mathbf{R}\times\mathbf{R}} \overline{f(x)}K^{1}(s,x;r,y)g(y)dx\,dy$$

$$= \int (f,d\nu^{\rm I}_{s;r}(X)g)\exp\Big[-i\int_r^s e\Phi(t,X(t))dt$$

$$+\, i\int_r^s eA(t,X(t))dX(t)\Big], \qquad (2.3)$$

with (f,g) in $\mathscr{S}(\mathbf{R};(\mathbf{C}^2)')\times\mathscr{S}(\mathbf{R};\mathbf{C}^2)$. The support of $\nu^{\rm I}_{s;r}$ is on the set of the Lipschitz continuous paths X: $|r,s|\rightarrow\mathbf{R}$ satisfying for each a,b with $r\leqslant a<b\leqslant s$ when $r<s$ or $r\geqslant a>b\geqslant s$ when $r>s$,

$$|X(b)-X(a)|\leqslant|b-a|\,,$$

$$[\,|X(t)-X(r)|=|t-r|,\quad \text{for } t\in|r,s|,\quad \text{in case } m=0]\,. \qquad (2.4)$$

(2) There exists a unique $\mathscr{S}'(\mathbf{R}^2\times\mathbf{R}^2;M_2(\mathbf{C}))$-valued countably additive measure $\nu^{\rm II}_{s;r}$ on the Banach space $C(|r,s|;\mathbf{R}^2)$ of the two-dimensional continuous paths $\mathbf{X}$: $|r,s|\rightarrow\mathbf{R}^2$, $\mathbf{X}(\tau)=(X^0(\tau),X^1(\tau))$, such that for every continuous $\mathbf{A}(\mathbf{x})=(A_0(\mathbf{x}),A_1(\mathbf{x}))$,

$$(f,e^{-i(s-r)H}g) = \int\int_{\mathbf{R}^2\times\mathbf{R}^2}\overline{f(x)}K^{\rm II}(s,\mathbf{x};r,\mathbf{y})g(\mathbf{y})d\mathbf{x}\,dy$$

$$= \int (f,d\nu^{\rm II}_{s;r}(\mathbf{X})g)$$

$$\times\exp\Big[-i\sum_{\rho=0}^{1}\int_r^s eA_\rho(\mathbf{X}(\tau))dX^\rho(\tau)\Big], \qquad (2.5)$$

with (f,g) in $\mathscr{S}(\mathbf{R}^2;(\mathbf{C}^2)')\times\mathscr{S}(\mathbf{R}^2;\mathbf{C}^2)$. The support of $\nu^{\rm II}_{s;r}$ is on the set of the Lipschitz continuous paths $\mathbf{X}$: $|r,s|\rightarrow\mathbf{R}^2$ satisfying for each a,b with $r\leqslant a<b\leqslant s$ when $r<s$ or $r\geqslant a>b\geqslant s$ when $r>s$,

$$X^0(b)-X^0(a)=b-a,\quad |X^1(b)-X^1(a)|\leqslant|b-a|\,,$$

$$[\,|X^1(t)-X^1(r)|=|t-r|,$$

$$\text{for } t\in|r,s|,\quad \text{in case } m=0]\,. \qquad (2.6)$$

In Ref. 4 we have improved on the support property of these path space measures $\nu^{\rm I}_{s;r}$ and $\nu^{\rm II}_{s;r}$ when the mass m is not zero, and, in fact, proved that almost every path with respect to them has a differential coefficient that is constant and exactly equal to plus or minus the light velocity except on a closed subset of time of Lebesgue measure zero.

In the present paper we want to prove the following theorem, which gives the ultimate result on the support property of the path space measures $\nu^{\rm I}_{s;r}$ and $\nu^{\rm II}_{s;r}$.

Theorem: (1) When $m>0$, the measure $\nu^{\rm I}_{s;r}$ is concentrated on the set of those Lipschitz continuous paths X: $|r,s|\rightarrow\mathbf{R}$ which satisfy

for some finite partition,

$r=t_0\lessgtr t_1\lessgtr\cdots\lessgtr t_k=s$ of $|r,s|$, depending on X,

$$X(t)-X(r)=\sum_{i=1}^{j-1}(-1)^i(t_i-t_{i-1})$$

$$+(-1)^j(t-t_{j-1})\,,$$

or $\qquad (2.7)$

$$X(t)-X(r)=\sum_{i=1}^{j-1}(-1)^{i-1}(t_i-t_{i-1})$$

$$+(-1)^{j-1}(t-t_{j-1})\,,$$

for $t\in|t_{j-1},t_j|$, $1\leqslant j\leqslant k$.

(2) When $m>0$, the measure $\nu^{\rm II}_{s;r}$ is concentrated on the set of those Lipschitz continuous paths $\mathbf{X}$: $|r,s|\rightarrow\mathbf{R}^2$ that satisfy

$X^0(t)-X^0(r)=t-r,\quad$ for $t\in|r,s|$,

and, for some finite partition, $r=t_0\lessgtr t_1\lessgtr\cdots\lessgtr t_k=s$ of $|r,s|$, depending on X,

$$X^1(t)-X^1(r)=\sum_{i=1}^{j-1}(-1)^i(t_i-t_{i-1})$$

$$+(-1)^j(t-t_{j-1})\,,$$

or $\qquad (2.8)$

$$X^1(t)-X^1(r)=\sum_{i=1}^{j-1}(-1)^{i-1}(t_i-t_{i-1})$$

$$+(-1)^{j-1}(t-t_{j-1})\,,$$

for $t\in|t_{j-1},t_j|$, $1\leqslant j\leqslant k$.

Remark 1: The support properties of $\nu^{\rm I}_{s;r}$ and $\nu^{\rm II}_{s;r}$ in the Theorem tell us the nature of the *Zitterbewegung*[8] of the Dirac particle. The motion described by a path satisfying (2.7) or (2.8) is such that the velocity is, in magnitude, equal to 1, the light velocity, in every finite time interval except for finitely many instants of time where the velocity alters its sign. Here the role the mass m plays is not to render the magnitude of the velocity smaller than 1, but to change the direction of motion of the particle time after time.

Remark 2: Feynman and Hibbs[5] give briefly a cryptic description of the fundamental solution of the Cauchy problem for the*free* Dirac equation in two space-time dimensions (see also Riazanov[6] and Rosen[7]).

Remark 3: Recently Blanchard *et al.*[9] have derived a corresponding support property from a basic property of Poisson processes. See also Gaveau *et al.*,[10] Gaveau,[11] Jacobson,[12] and De Angelis *et al.*[13] Our proof is, however, a direct and analytic derivation from the path space measures, not passing through a Poisson process.

III. PROOF OF THEOREM

We shall only prove the first part of the Theorem here. The second part will be shown similarly. Without loss of generality, we may assume that $r=0$ and $s>0$. Before proving the support property of the measure $\nu^{\rm I}_{s;0}$ as in the Theorem, we sketch the way to construct $\nu^{\rm I}_{s;0}$.

Consider the Cauchy problem for the free equation to (2.1),

$$\partial_t\phi(t,x)=[-\alpha\,\partial_x-im\beta\,]\phi(t,x),\quad t\in\mathbf{R},\quad x\in\mathbf{R}\,, \qquad (3.1)$$

with initial data $\phi(0,x)=g(x)$. Let $K^{\rm I}_0(s,x)$ be the fundamental solution:

$$\phi(s,x)=(e^{-s(\alpha\,\partial_x+im\beta)}g)(x)=\int_{\mathbf{R}}K^{\rm I}_0(s,x-y)g(y)dy\,. \qquad (3.2)$$

It is given by

$$K_0^{\mathrm{I}}(s,x) = 2^{-1}[\partial_s - \alpha\,\partial_x - im\beta\,] \times (J_0(m(s^2-x^2)^{1/2})\theta(s-|x|)), \quad (3.3)$$

where $J_0(t)$ is the Bessel function of order zero, and $\theta(t)$ the Heaviside function $\theta(t) = 1$ for $t>0$, $=0$ for $t<0$. Here $C_\infty(\mathbb{R};\mathbb{C}^2)$ denotes the Banach space of the $\mathbb{C}^2$-valued continuous functions in $\mathbb{R}$ that vanish at infinity. Using Eq. (3.3), we get the following lemma.

Lemma 1: The $e^{-t(\alpha\,\partial_x + im\beta)}$ is a continuous linear operator of $C_\infty(\mathbb{R};\mathbb{C}^2)$ into itself, and satisfies

$$\|Ne^{-t(\alpha\,\partial_x+im\beta)}g\|_\infty < e^{m|t|}\|Ng\|_\infty\,, \quad (3.4)$$

for g in $C_\infty(\mathbb{R};\mathbb{C}^2)$. Here N is a unitary matrix satisfying

$$N\alpha N^{-1} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.$$

For each fixed $s>0$ let $\mathscr{X}_{s,0} = \Pi_{[0,s]}\dot{\mathbb{R}} = (\dot{\mathbb{R}})^{[0,s]}$ be the product of the uncountably many copies of $\dot{\mathbb{R}}$, where $\dot{\mathbb{R}} = \mathbb{R}\cup\{\infty\}$ is the one-point compactification of $\mathbb{R}$. By the Tychonoff theorem[14] $\mathscr{X}_{s,0}$ is a compact Hausdorff space in the product topology. It may be regarded as the space of all paths X: $[0,s]\to\dot{\mathbb{R}}$, possibly discontinuous and possibly passing through infinity. Let $C(\mathscr{X}_{s,0})$ be the Banach space of the complex-valued continuous functions on $\mathscr{X}_{s,0}$ and $C_{\mathrm{fin}}(\mathscr{X}_{s,0})$ the subspace of those Ψ in $C(\mathscr{X}_{s,0})$ for which there exists a finite partition:

$$0=s_0<s_1<\cdots<s_n=s\,, \quad (3.5)$$

of the interval $[0,s]$ and a complex-valued bounded continuous function $F(x_0,x_1,...,x_n)$ on $(\dot{\mathbb{R}})^{n+1}$ such that

$$\Psi(X) = F(X(s_0),X(s_1),...,X(s_n))\,. \quad (3.6)$$

Define, for each fixed $s>0$, a functional $L_{s,0}(\Psi;f,g)$ linear in $\Psi\in C_{\mathrm{fin}}(\mathscr{X}_{s,0})$ and sesquilinear in $(f,g) \in\mathscr{S}(\mathbb{R};(\mathbb{C}^2)')\times\mathscr{S}(\mathbb{R};\mathbb{C}^2)$ by

$$\begin{aligned} L_{s,0}(\Psi;f,g) = \int_{\mathbb{R}} dx_0\cdots\int_{\mathbb{R}} dx_n\ &\overline{f(x_n)} \\ &\times K_0^{\mathrm{I}}(s_n-s_{n-1},x_n-x_{n-1}) \\ &\times K_0^{\mathrm{I}}(s_{n-1}-s_{n-2},x_{n-1}-x_{n-2}) \\ &\times\cdots\times K_0^{\mathrm{I}}(s_1-s_0,x_1-x_0) \\ &\times F(x_0,x_1,...,x_n)g(x_0)\,. \end{aligned} \quad (3.7)$$

Then we can show the following lemma by successively using Lemma 1.

Lemma 2: (1) For each fixed (f,g) in $\mathscr{S}(\mathbb{R};(\mathbb{C}^2)')\times\mathscr{S}(\mathbb{R};\mathbb{C}^2)$, $L_{s,0}(\Psi;f,g)$ is well defined on $C_{\mathrm{fin}}(\mathscr{X}_{s,0})$; it is independent of the choice of F corresponding to Ψ.

(2) The following inequality holds:

$$|L_{s,0}(\Psi;f,g)| < ce^{m|s|}\|\Psi\|\ \|f\|_1\|g\|_\infty\,, \quad (3.8)$$

for every Ψ in $C_{\mathrm{fin}}(\mathscr{X}_{s,0})$ and every pair (f,g) in $\mathscr{S}(\mathbb{R};(\mathbb{C}^2)')\times\mathscr{S}(\mathbb{R};\mathbb{C}^2)$, with $c = |N|\,|N^{-1}|<2$. Here the norm of a 2×2 matrix $M = (M_{jk})$ is defined by

$$|M| = \max_{1<j<2}\sum_{k=1}^{2}|M_{jk}|\,.$$

The L^1 and L^∞ norms are denoted by $\|\cdot\|_1$ and $\|\cdot\|_\infty$, respectively.

Since $C_{\mathrm{fin}}(\mathscr{X}_{s,0})$ is dense in $C(\mathscr{X}_{s,0})$ by the Stone–Weierstrass theorem,[14] the inequality (3.8) holds also for $\Psi\in C(\mathscr{X}_{s,0})$. Then in virtue of the Riesz-type representation theorem,[15] there exists a unique $\mathscr{S}'(\mathbb{R}\times\mathbb{R};M_2(\mathbb{C}))$-valued Borel measure $\nu_{0;s}^{\mathrm{I}}$ on $\mathscr{X}_{s,0}$ such that

$$\int_{\mathscr{X}_{s,0}}(f,d\nu_{0;s}^{\mathrm{I}}(X)g)\Psi(X) = L_{s,0}(\Psi;f,g)\,.$$

We shall now see the measure $\nu_{s;0}^{\mathrm{I}}$ has the support property described in part (1) of the Theorem.

First we explain an outline of the following argument. Rewrite the functional $L_{s,0}(\Psi;f,g)$ in (3.7) as

$$\begin{aligned} L_{s,0}(\Psi;f,g) = (f(x_n),&N^{-1}e^{(s_n-s_{n-1})C}\cdots e^{(s_1-s_0)C} \\ &\times F(x_0,...,x_n)Ng(x_0))\,, \end{aligned} \quad (3.7')$$

where $C = -N[\alpha\,\partial_x + im\beta]N^{-1}$ with the unitary matrix N in Lemma 1, and for $j=1,2,...,n$, the operator $e^{(s_j-s_{j-1})C}$ maps the function of x_{j-1} into that of x_j. Put

$$A = -N\alpha N^{-1}\,\partial_x \quad\text{and}\quad B = -imN\beta N^{-1}\,,$$

so that

$$e^{tC} = Ne^{-t(\alpha\,\partial_x+im\beta)}N^{-1} = e^{t(A+B)}\,.$$

By iterating the formula

$$e^{tC} = e^{tA} + \int_0^t d\tau_1\,e^{(t-\tau_1)A}Be^{\tau_1 C}\,, \quad (3.9)$$

we make the Taylor expansion of e^{tC} in m:

$$\begin{aligned} e^{tC} = \sum_{k=0}^{\infty}\int_0^\infty d\tau_1\cdots\int_0^\infty d\tau_k\ &\theta\Big(t-\sum_{i=1}^{k}\tau_i\Big) \\ \times\exp\Big[\Big(t-&\sum_{i=1}^{k}\tau_i\Big)A\Big]B\exp(\tau_k A)B \\ \times\exp(\tau_{k-1}A)&\cdots B\exp(\tau_1 A)\equiv\sum_{k=0}^{\infty}S_t^k\,. \end{aligned} \quad (3.10)$$

We substitute (3.10) into (3.7′) to get

$$\begin{aligned} L_{s,0}(\Psi;f,g) &= \sum_{k=0}^{\infty}\sum_{\substack{k_1+\cdots+k_n=k \\ k_j>0}}(f(x_n),N^{-1}S^{k_n}_{s_n-s_{n-1}} \\ &\qquad\times\cdots\times S^{k_1}_{s_1-s_0}F(x_0,...,x_n)Ng(x_0)) \\ &\equiv\sum_{k=0}^{\infty}L^k_{s,0}(\Psi;f,g)\,, \end{aligned}$$

where for $j=1,2,...,n$, the operator $S^{k_j}_{s_j-s_{j-1}}$ maps the function of x_{j-1} to that of x_j. Next we can show, for each functional $L^k_{s,0}(\Psi;f,g)$, a lemma analogous to Lemma 2, which will yield a Borel measure $\nu^k_{s;0}$ on $\mathscr{X}_{s,0}$ with

$$\int_{\mathscr{X}_{s,0}}(f,d\nu^k_{s;0}(X)g)\Psi(X) = L^k_{s,0}(\Psi;f,g)\,,$$

so that

$$\nu^{\mathrm{I}}_{s;0} = \sum_{k=0}^{\infty}\nu^k_{s;0}\,.$$

Finally we show each $\nu^k_{s;0}$ is concentrated on the zigzag paths of k straight segments and hence conclude $\nu^{\mathrm{I}}_{s;0}$ possesses the desired support property.

Now we carry out the procedure described above. Set

$$S^0_t = e^{tA}, \quad S^{0\to}_t = e^{tC} = e^{t(A+B)}, \tag{3.11a}$$

and

$$\begin{aligned} S^k_t = {} & \int_0^\infty d\tau_1 \cdots \int_0^\infty d\tau_k\, \theta\Big(t - \sum_{i=1}^k \tau_i\Big) \\ & \times \exp\Big[\Big(t - \sum_{i=1}^k \tau_i\Big)A\Big] B \exp(\tau_k A) B \\ & \times \exp(\tau_{k-1}A) \cdots B \exp(\tau_1 A), \\ S^{k\to}_t = {} & \int_0^\infty d\tau_1 \cdots \int_0^\infty d\tau_k\, \theta\Big(t - \sum_{i=1}^k \tau_i\Big) \\ & \times \exp\Big[\Big(t - \sum_{i=1}^k \tau_i\Big)A\Big] B \exp(\tau_k A) \\ & \times \cdots \times B \exp(\tau_2 A) B \exp(\tau_1 C), \end{aligned} \tag{3.11b}$$

for $k \geqslant 1$.

Then we have the following lemma.

Lemma 3:

(1) $S^{0\to}_t = \sum_{k=0}^N S^k_t + S^{N+1\to}_t, \quad N = 0,1,2,\ldots.$

(2) S^k_t and $S^{k\to}_t$, $k = 0,1,\ldots$, are bounded linear operators of $C_\infty(\mathbb{R};\mathbb{C}^2)$ into itself:

$$\|S^k_t\| \leqslant (k!)^{-1}(mt)^k, \quad \|S^{k\to}_t\| \leqslant (k!)^{-1}(mt)^k e^{mt}.$$

Proof: By iteration of (3.9), we get (1). The statement (2) is a direct consequence of definition (3.11) and the estimates

$$\|e^{tA}\| \leqslant 1, \quad \|B\| = m, \quad \|e^{tC}\| \leqslant e^{mt}.$$

The estimate $\|e^{tC}\| \leqslant e^{mt}$ is nothing but (3.4). Since $N\beta N^{-1}$ anticommutes with $N\alpha N^{-1}$ and $N\alpha N^{-1}$ is diagonal, we have

$$(N\beta N^{-1})_{11} = (N\beta N^{-1})_{22} = 0,$$
$$|(N\beta N^{-1})_{12}| = |(N\beta N^{-1})_{21}| = 1$$

and hence $\|B\| = m$. Notice that e^{tA} operates on

$$\begin{pmatrix} \varphi_1 \\ \varphi_2 \end{pmatrix} \in C_\infty(\mathbb{R};\mathbb{C}^2)$$

according to

$$\left(e^{tA}\begin{pmatrix} \varphi_1 \\ \varphi_2 \end{pmatrix}\right)(x) = \begin{pmatrix} \varphi_1(x-t) \\ \varphi_2(x+t) \end{pmatrix}, \tag{3.12}$$

so that we get $\|e^{tA}\| = 1$. □

For $\Psi \in C_{\mathrm{fin}}(\mathscr{X}_{s,0})$ represented as (3.6) with a continuous function $F(x_0,x_1,\ldots,x_n)$ on $(\dot{\mathbb{R}})^{n+1}$, we introduce a sequence

$$\{F^{(l)}_{K_1,\ldots,K_l;x_{l+1},\ldots,x_n}\}_{l=1}^n$$

of $\mathbb{C}^2$-valued functions on $\mathbb{R}$ with parameters. Set

$$F^{(0)}_{x_1,\ldots,x_n}(x) = Ng(x)F(x,x_1,\ldots,x_n), \tag{3.13a}$$

and with S^k_t and $S^{k\to}_t$ in (3.11),

$$\begin{aligned} & F^{(l)}_{K_1,\ldots,K_l;x_{l+1},\ldots,x_n}(x) \\ & \quad = (S^{K_l}_{s_l - s_{l-1}} F^{(l-1)}_{K_1,\ldots,K_{l-1};x,x_{l+1},\ldots,x_n})(x), \end{aligned} \tag{3.13b}$$

for $l = 1,2,\ldots,n-1$, and

$$F^{(n)}_{K_1,\ldots,K_n}(x) = (S^{K_n}_{s_n - s_{n-1}} F^{(n-1)}_{K_1,\ldots,K_{n-1};x})(x). \tag{3.13c}$$

Here K_l is k_l or $k_l\to$ with k_l a non-negative integer. For $1 \leqslant l \leqslant n-1$,

$$F^{(l-1)}_{K_1,\ldots,K_{l-1};x,x_{l+1},\ldots,x_n}(y)$$

in (3.13b) is in $C_\infty(\mathbb{R};\mathbb{C}^2)$ as a function of y.

For each $k \geqslant 0$, define the functionals $L^k_{s,0}(\Psi;f,g)$ and $L^{k\to}_{s,0}(\Psi;f,g)$, which are linear in $\Psi \in C_{\mathrm{fin}}(\mathscr{X}_{s,0})$ and sesquilinear in $(f,g) \in \mathscr{S}(\mathbb{R};(\mathbb{C}^2)') \times \mathscr{S}(\mathbb{R};\mathbb{C}^2)$, by

$$\begin{aligned} L^0_{s,0}(\Psi;f,g) & \equiv \int_{\mathbf{R}} \overline{f(x)} N^{-1} F^{(n)}_{\underbrace{0,\ldots,0}_{n}}(x)\,dx, \\ L^{0\to}_{s,0}(\Psi;f,g) & \equiv \int_{\mathbf{R}} \overline{f(x)} N^{-1} F^{(n)}_{\underbrace{0\to,\ldots,0\to}_{n}}(x)\,dx, \end{aligned} \tag{3.14a}$$

and

$$\begin{aligned} L^k_{s,0}(\Psi;f,g) \equiv {} & \sum_{\Sigma_{i=1}^n k_i = k,\ k_1,\ldots,k_n \geqslant 0} \int_{\mathbf{R}} \overline{f(x)} N^{-1} \\ & \times F^{(n)}_{k_1,\ldots,k_n}(x)\,dx, \end{aligned} \tag{3.14b}$$

$$\begin{aligned} L^{k\to}_{s,0}(\Psi;f,g) \equiv {} & \sum_{p=1}^n \sum_{\substack{\Sigma_{i=p}^n k_i = k \\ k_p \geqslant 1,\ k_{p+1},\ldots,k_n \geqslant 0}} \int_{\mathbf{R}} \overline{f(x)} N^{-1} \\ & \times F^{(n)}_{\underbrace{0\to,\ldots,0\to}_{p-1},k_p\to,k_{p+1},\ldots,k_n}(x)\,dx, \end{aligned} \tag{3.14c}$$

for $k \geqslant 1$. Note that $L^{0\to}_{s,0}(\Psi;f,g)$ in (3.14a) is nothing but $L_{s,0}(\Psi;f,g)$ in (3.7).

Then the following lemma holds.

Lemma 4: (1) For each fixed $(f,g) \in \mathscr{S}(\mathbb{R};(\mathbb{C}^2)') \times \mathscr{S}(\mathbb{R};\mathbb{C}^2)$ and each $k \geqslant 0$, $L^k_{s,0}(\Psi;f,g)$ and $L^{k\to}_{s,0}(\Psi;f,g)$ are well defined on $C_{\mathrm{fin}}(\mathscr{X}_{s,0})$; they are independent of the choice of F corresponding to Ψ.

(2) The following inequalities hold:

$$|L^k_{s,0}(\Psi;f,g)| \leqslant c(k!)^{-1}(ms)^k \|\Psi\|\, \|f\|_1 \|g\|_\infty,$$
$$|L^{k\to}_{s,0}(\Psi;f,g)| \leqslant c(k!)^{-1}(ms)^k e^{ms} \|\Psi\|\, \|f\|_1 \|g\|_\infty,$$

with a constant $c \leqslant 2$.

(3)
$$\begin{aligned} L^{0\to}_{s,0}(\Psi;f,g) = {} & \sum_{l=0}^k L^l_{s,0}(\Psi;f,g) \\ & + L^{k+1\to}_{s,0}(\Psi;f,g), \end{aligned}$$
$$L^{0\to}_{s,0}(\Psi;f,g) = \sum_{l=0}^\infty L^l_{s,0}(\Psi;f,g).$$

Proof: The statement (1) follows from the identities

$$\sum_{l=0}^k S^l_{\tau_1} S^{k-l}_{\tau_2} = S^k_{\tau_1+\tau_2}$$

and

$$\sum_{l=0}^{k-1} S^l_{\tau_1} S^{k-l\to}_{\tau_2} + S^{k\to}_{\tau_1} S^{0\to}_{\tau_2} = S^{k\to}_{\tau_1+\tau_2}, \quad \tau_1,\tau_2 > 0,$$

which can be derived from the definition (3.11). To prove (2), we first note that by induction and continuity of the operator S^k_t in Lemma 3,

$$F^{(l)}_{k_1,\ldots,k_l;x_{l+1},\ldots,x_n}$$

is in $C_\infty(\mathbb{R};\mathbb{C}^2)$ and continuous as a map on the parameter space $\mathbb{R}^{n-1}$ into $C_\infty(\mathbb{R};\mathbb{C}^2)$, for fixed $l = 1,...,n$. Recalling the definitions (3.13b) and (3.13c), we obtain, by iterative application of Lemma 3 (2),

$$\begin{aligned}\|F^{(n)}_{k_1,...,k_n}\|_\infty &= \sup_{x\in\mathbf{R},\, j=1,2} |(S^{k_n}_{s_n - s_{n-1}} F^{(n-1)}_{k_1,...,k_{n-1};x})_j(x)| \\ &\leqslant \sup_{x\in\mathbf{R}} \|S^{k_n}_{s_n - s_{n-1}} F^{(n-1)}_{k_1,...,k_{n-1};x}\|_\infty \\ &\leqslant (m(s_n - s_{n-1}))^{k_n}(k_n!)^{-1} \\ &\quad \times \sup_{x\in\mathbf{R}} \|F^{(n-1)}_{k_1,...,k_{n-1};x}\|_\infty \leqslant \cdots \\ &\leqslant (m(s_1 - s_0))^{k_1}\cdots(m(s_n - s_{n-1}))^{k_n} \\ &\quad \times (k_1!)^{-1}\cdots(k_n!)^{-1} \\ &\quad \times \sup_{x_1,...x_n\in\mathbf{R}} \|F^{(0)}_{x_1,...,x_n}\|_\infty .\end{aligned}$$

This estimate together with definitions (3.14b) and (3.13a) yields the first inequality in (2). We get similarly the second one in (2). The first equality in (3) is a direct consequence of Lemma 3 (1). Note that $|L^{k+1\rightarrow}_{s,0}(\Psi;f,g)| \rightarrow 0$ as $k \rightarrow \infty$ by (2), then the second equality in (3) holds. □

The consequence of Lemma 4 is the following proposition.

Proposition 5: For each $k \geqslant 0$, there exist unique complex-valued countably additive regular measures $\nu^k_{s,f;0,g}$ and $\nu^{k\rightarrow}_{s,f;0,g}$ on the Borel sets in $\mathscr{X}_{s,0}$ such that for each Ψ in $C(\mathscr{X}_{s,0})$,

$$L^k_{s,0}(\Psi;f,g) = \int_{\mathscr{X}_{s,0}} d\nu^k_{s,f;0,g}(X)\Psi(X) ,$$

$$L^{k\rightarrow}_{s,0}(\Psi;f,g) = \int_{\mathscr{X}_{s,0}} d\nu^{k\rightarrow}_{s,f;0,g}(X)\Psi(X) .$$

Moreover the following equality holds for every Borel set E in $\mathscr{X}_{s,0}$:

$$\nu^{0\rightarrow}_{s,f;0,g}(E) = \nu^{\mathrm{I}}_{s,f;0,g}(E) = \sum_{k=0}^{\infty} \nu^k_{s,f;0,g}(E) ,$$

where the series in the last member is absolutely convergent. Therefore if, for each $k \geqslant 0$, the measure $\nu^k_{s,f;0,g}$ is concentrated on a Borel subset E_k of $\mathscr{X}_{s,0}$, then the measure $\nu^{\mathrm{I}}_{s,f;0,g}$ is concentrated on the Borel subset $\cup_{k=0}^{\infty} E_k$.

Proof: The first half of the proposition follows from Lemma 4 (1), 4 (2), and the Riesz representation theorem. Next this and Lemma 4 (3) yield

$$\nu^{\mathrm{I}}_{s,f;0,g} = \sum_{k=0}^{N} \nu^k_{s,f;0,g} + \nu^{N+1\rightarrow}_{s,f;0,g} .$$

Further, by Lemma 4 (2), we get

$$|\nu^{N+1\rightarrow}_{s,f;0,g}(E)| \leqslant c((N+1)!)^{-1}(ms)^{N+1}e^{ms}\|f\|_1\|g\|_\infty ,$$

which converges to zero as $N \rightarrow \infty$. This proves the second half of the proposition. □

Our next task is to see the support property of $\nu^{\mathrm{I}}_{s;0}$. We shall show in Proposition 7 below that, for each $k \geqslant 0$, the measure $\nu^k_{s,f;0,g}$ is concentrated on the set of the Lipschitz continuous paths X: $[0,s] \rightarrow \mathbb{R}$ satisfying, for some k-partition $0 = t_0 < t_1 < \cdots < t_k = s$ of the interval $[0,s]$, depending on X,

$$\begin{aligned}X(t) - X(0) = \sum_{i=1}^{j-1} (-1)^i(t_i - t_{i-1}) \\ + (-1)^j(t - t_{j-1})\end{aligned}$$

or

$$\begin{aligned}X(t) - X(0) = \sum_{i=1}^{j-1} (-1)^{i-1}(t_i - t_{i-1}) \\ + (-1)^{j-1}(t - t_{j-1}) ,\end{aligned}$$

for $t_{j-1} \leqslant t \leqslant t_j$, $1 \leqslant j \leqslant k$.

For each $k \geqslant 1$, let Δ^k be the open k-simplex

$$\Delta^k = \left\{ (\tau_1,...,\tau_k)\in\mathbf{R}^k \;\middle|\; \sum_{i=1}^{k} \tau_i < s \text{ and } \tau_1,...,\tau_k > 0 \right\} ,$$

and φ^k_1 and φ^k_2 the maps from $\mathbf{R}\times\Delta^k$ into $\mathscr{X}_{s,0}$ defined by

$$\begin{aligned}&\varphi^k_j(x,\tau_1,...,\tau_k)(t) \\ &= x + (-1)^j \\ &\quad \times \left[\sum_{l=1}^{N_t} (-1)^l\tau_l + (-1)^{N_t+1}\left(t - \sum_{l=1}^{N_t} \tau_l\right)\right] ,\end{aligned} \qquad (3.15)$$

for $x\in\mathbf{R}$, $(\tau_1,...,\tau_k)\in\Delta^k$, $t\in[0,s]$, and $j = 1,2$, where N_t is the t-dependent integer satisfying

$$\sum_{l=1}^{N_t} \tau_l \leqslant t < \sum_{l=1}^{N_t+1} \tau_l .$$

In (3.15), the value $\varphi^k_j(x,\tau_1,...,\tau_k)$ is a function: $[0,s] \rightarrow \mathbf{R}$ and so belongs to $\mathscr{X}_{s,0}$.

For $k = 0$, we understand Δ^0 to be the set of one point and identify $\mathbf{R}\times\Delta^0$ with $\mathbf{R}$. Let φ^0_1 and φ^0_2 be the maps from $\mathbf{R} = \mathbf{R}\times\Delta^0$ into $\mathscr{X}_{s,0}$ defined by

$$\varphi^0_j(x)(t) = x + (-1)^j t, \quad j = 1,2 .$$

Then the maps φ^k_j, $k \geqslant 0$, $j = 1,2$, have the following properties.

Lemma 6: (1) For each $k \geqslant 0$ and $j = 1,2$, φ^k_j is continuous and Borel measurable.

(2) $\varphi^k_j(\mathbf{R}\times\Delta^k)$ is an F_σ set.

Proof: Statement (1) is obvious.

(2) $\mathbf{R}\times\Delta^k$ is expressed as a countable union $\cup_{n=1}^{\infty} K_n$ of compact sets K_n. By the continuity of φ^k_j, each $\varphi^k_j(K_n)$ is compact in $\mathscr{X}_{s,0}$ and hence closed, so $\varphi^k_j(\mathbf{R}\times\Delta^k)$ is an F_σ set. □

For each $k \geqslant 0$ and $j = 1,2$, define the complex-valued regular Borel measure $\mu^{k,j}_{s,f;0,g}$ on $\mathbf{R}\times\Delta^k$ by

$$\mu^{k,j}_{s,f;0,g}(E) = \int_E \left(\overline{f\left(x + (-1)^j\left[\sum_{l=1}^{k} (-1)^l\tau_l + (-1)^{k+1}\left(s - \sum_{l=1}^{k} \tau_l\right)\right]\right)} N^{-1}B^k \right)_j (Ng(x))_j \, dx\, d\tau_1\cdots d\tau_k , \qquad (3.16)$$

where E is a Borel set in $\mathbb{R}\times\Delta^k$.

Proposition 7: For each $(f,g)\in\mathscr{S}(\mathbb{R};(\mathbb{C}^2)')\times\mathscr{S}(\mathbb{R};\mathbb{C}^2)$, $k\geqslant 0$ and $s>0$,

$$\nu^k_{s,f;0,g}=\sum_{j=1}^{2}\varphi_j^k\cdot\mu^{k,j}_{s,f;0,g}\,.$$

Here $\varphi_j^k\cdot\mu^{k,j}_{s,f;0,g}$ is the image measure[16] on $\mathscr{X}_{s,0}$ induced from the measure $\mu^{k,j}_{s,f;0,g}$ on $\mathbb{R}\times\Delta^k$ by the map φ_j^k.

Before showing Proposition 7, we see first what is a consequence of the proposition. The measure $\varphi_j^k\cdot\mu^{k,j}_{s,f;0,g}$ is concentrated on the set $\varphi_j^k(\mathbb{R}\times\Delta^k)$ and thus the measure $\nu^k_{s,f;0,g}$ on the set $\cup_{j=1}^2\varphi_j^k(\mathbb{R}\times\Delta^k)$. It follows by Proposition 5 that the measure $\nu^{\mathrm{I}}_{s,f;0,g}$ is concentrated on the set

$$\cup_{k=0}^{\infty}\cup_{j=1}^{2}\varphi_j^k(\mathbb{R}\times\Delta^k)\,,$$

which is a Borel set, in fact, an F_σ set, by Lemma 6. Then the $\mathscr{S}'(\mathbb{R}\times\mathbb{R};M_2(\mathbb{C}))$-valued measure $\nu^{\mathrm{I}}_{s;0}$ is also concentrated on the set

$$\cup_{k=0}^{\infty}\cup_{j=1}^{2}\varphi_j^k(\mathbb{R}\times\Delta^k)\,,$$

since this set is independent of (f,g) in $\mathscr{S}(\mathbb{R};(\mathbb{C}^2)')\times\mathscr{S}(\mathbb{R};\mathbb{C}^2)$. It is obvious that every X in

$$\cup_{k=0}^{\infty}\cup_{j=1}^{2}\varphi_j^k(\mathbb{R}\times\Delta^k)$$

satisfies the condition (2.7) in Theorem. In case $m=0$, we have $\nu^{\mathrm{I}}_{s,f;0,g}=\nu^{0}_{s,f;0,g}$. By Proposition 7 it is concentrated on the set of the straight segments X: $[0,s]\to\mathbb{R}$ with $X(t)=X(0)+t$ or $X(t)=X(0)-t$, $0\leqslant t\leqslant s$, and so is $\nu^{\mathrm{I}}_{s;0}$, similarly. Thus we have seen Proposition 7 yields the desired support property of $\nu^{\mathrm{I}}_{s;0}$.

Proof of Proposition 7: It is enough to show that the equality

$$\int_{\mathscr{X}_{s,0}}d\nu^k_{s,f;0,g}(X)\Psi(X)$$
$$=\sum_{j=1}^{2}\int_{\mathbb{R}\times\Delta^k}d\mu^{k,j}_{s,f;0,g}(\zeta)\Psi(\varphi_j^k(\zeta)),$$
$$\zeta=(x,\tau_1,\dots,\tau_k)\,,$$

holds for every $\Psi\in C_{\mathrm{fin}}(\mathscr{X}_{s,0})$. Here we only prove this equality for $k=2$. The proof will be still complicated for general Ψ in $C_{\mathrm{fin}}(\mathscr{X}_{s,0})$. So we only see it when Ψ is represented as $\Psi(X)=F(X(0),X(s_1),X(s))$ with a partition $0=s_0<s_1<s_2=s$ of the interval $[0,s]$ and a bounded continuous function F on $(\dot{\mathbb{R}})^3$, i.e., (3.5) and (3.6) with $n=2$. We can similarly prove the general case.

Recall that

$$e^{\tau A}=\begin{pmatrix}e^{-\tau\partial_x}&0\\0&e^{-\tau\partial_x}\end{pmatrix},\quad B=\begin{pmatrix}0&B_{12}\\B_{21}&0\end{pmatrix},$$

with $|B_{12}|=|B_{21}|=m$.

By Proposition 5 and definition (3.14), we have

$$\int_{\mathscr{X}_{s,0}}d\nu^2_{s,f;0,g}(X)\Psi(X)$$
$$=L^2_{s,0}(\Psi;f,g)$$
$$=\int_{\mathbf{R}}\overline{f(x)}N^{-1}\{F^{(2)}_{0,2}(x)+F^{(2)}_{1,1}(x)+F^{(2)}_{2,0}(x)\}dx\,. \tag{3.17}$$

Using (3.13), (3.11), and the anticommutativity of A and B, we get

$$F^{(2)}_{1,1}(x)=(S^1_{s-s_1}F^{(1)}_{1;x})(x)$$
$$=B\int_0^\infty d\tau_2\,\theta(s-s_1-\tau_2)$$
$$\times\,(e^{(-s+s_1+2\tau_2)A}F^{(1)}_{1;x})(x)$$
$$=B\int_0^\infty d\tau_2\,\theta(s-s_1-\tau_2)$$
$$\times\begin{pmatrix}(F^{(1)}_{1;x})_1(x+s-s_1-2\tau_2)\\(F^{(1)}_{1;x})_2(x-s+s_1+2\tau_2)\end{pmatrix}.$$

Here

$$F^{(1)}_{1;x}(y)={}^t((F^{(1)}_{1;x})_1(y),(F^{(1)}_{1;x})_2(y))\,.$$

For these functions in the integrand of the last member of the above equation, we get

$$(F^{(1)}_{1;x})_1(x+s-s_1-2\tau_2)$$
$$=(S^1_{s_1}F^{(0)}_{x+s-s_1-2\tau_2,x})_1(x+s-s_1-2\tau_2)$$
$$=B_{12}\int_0^\infty d\tau_1\,\theta(s_1-\tau_1)$$
$$\times\,(e^{(-s_1+2\tau_1)A}F^{(0)}_{x+s-s_1-2\tau_2,x})_2(x+s-s_1-2\tau_2)$$
$$=B_{12}\int_0^\infty d\tau_1\,\theta(s_1-\tau_1)$$
$$\times\,(Ng)_2(x+s-2s_1-2\tau_2+2\tau_1)$$
$$\times\,F(x+s-2s_1-2\tau_2+2\tau_1,x+s-s_1-2\tau_2,x)\,,$$

and

$$(F^{(1)}_{1;x})_2(x-s+s_1+2\tau_2)$$
$$=B_{21}\int_0^\infty d\tau_1\,\theta(s_1-\tau_1)$$
$$\times\,(Ng)_1(x-s+2s_1+2\tau_2-2\tau_1)$$
$$\times\,F(x-s+2s_1+2\tau_2-2\tau_1,x-s+s_1+2\tau_2,x)\,.$$

Hence we have

$$F^{(2)}_{1,1}(x)=B^2\int_0^\infty d\tau_2\int_0^\infty d\tau_1\,\theta(s-s_1-\tau_2)\theta(s_1-\tau_1)$$
$$\times\begin{pmatrix}(Ng)_1(x-s+2s_1+2\tau_2-2\tau_1)F(x-s+2s_1+2\tau_2-2\tau_1,x-s+s_1+2\tau_2,x)\\(Ng)_2(x+s-2s_1-2\tau_2+2\tau_1)F(x+s-2s_1-2\tau_2+2\tau_1,x+s-s_1-2\tau_2,x)\end{pmatrix}$$
$$=B^2\int_0^\infty d\tau_2\int_0^\infty d\tau_1\,\theta(s-\tau_1-\tau_2)\theta(\tau_1+\tau_2-s_1)\theta(s_1-\tau_1) \tag{3.18}$$

$$\times\begin{pmatrix}(Ng)_1(x-s+2\tau_2)F(x-s+2\tau_2,x-s-s_1+2\tau_2+2\tau_1,x)\\(Ng)_2(x+s-2\tau_2)F(x+s-2\tau_2,x+s+s_1-2\tau_2-2\tau_1,x)\end{pmatrix},$$

where, in the second equality, we have first made the change of variable $\tau_2'=\tau_2+s_1-\tau_1$ and next written τ_2 again instead of τ_2'. Similarly we have

$$F_{0,2}^{(2)}(x)=B^2\int_0^\infty d\tau_2\int_0^\infty d\tau_1\,\theta(s-\tau_1-\tau_2)\theta(\tau_1-s_1)\begin{pmatrix}(Ng)_1(x-s+2\tau_2)F(x-s+2\tau_2,x-s+s_1+2\tau_2,x)\\(Ng)_2(x+s-2\tau_2)F(x+s-2\tau_2,x+s-s_1-2\tau_2,x)\end{pmatrix},\tag{3.19}$$

and

$$F_{2,0}^{(2)}(x)=B^2\int_0^\infty d\tau_2\int_0^\infty d\tau_1\,\theta(s_1-\tau_1-\tau_2)\begin{pmatrix}(Ng)_1(x-s+2\tau_2)F(x-s+2\tau_2,x-s+s_1,x)\\(Ng)_2(x+s-2\tau_2)F(x+s-2\tau_2,x+s-s_1,x)\end{pmatrix}.\tag{3.20}$$

Substituting (3.18)–(3.20) into (3.17), we get

$$\begin{aligned}&\int_{\mathscr{X}_{s,0}} d\nu^2_{s,f;0,g}(X)\Psi(X)\\&\quad=\int_{\mathbf{R}}dx\int_0^\infty d\tau_1\int_0^\infty d\tau_2(\overline{f(x+s-2\tau_2)}N^{-1}B^2)_1\theta(s-\tau_1-\tau_2)[\theta(\tau_1-s_1)(Ng)_1(x)F(x,x+s_1,x+s-2\tau_2)\\&\qquad+\theta(\tau_1+\tau_2-s_1)\theta(s_1-\tau_1)(Ng)_1(x)F(x,x-s_1+2\tau_1,x+s-2\tau_2)+\theta(s_1-\tau_1-\tau_2)(Ng)_1(x)\\&\qquad\times F(x,x+s_1-2\tau_2,x+s-2\tau_2)]+\int_{\mathbf{R}}dx\int_0^\infty d\tau_1\int_0^\infty d\tau_2(\overline{f(x-s+2\tau_2)}N^{-1}B^2)_2\theta(s-\tau_1-\tau_2)\\&\qquad\times[\theta(\tau_1-s_1)(Ng)_2(x)F(x,x-s_1,x-s+2\tau_2)+\theta(\tau_1+\tau_2-s_1)\theta(s_1-\tau_1)(Ng)_2(x)\\&\qquad\times F(x,x+s_1-2\tau_1,x-s+2\tau_2)+\theta(s_1-\tau_1-\tau_2)(Ng)_2(x)F(x,x-s_1+2\tau_2,x-s+2\tau_2)]\,.\end{aligned}\tag{3.21}$$

Here, on the right-hand side, we have first made the change of variables $x'=x-s+2\tau_2$ in the first term and $x''=x+s-2\tau_2$ in the second term, and next written x instead of x' and x''.

By the definition (3.15) of $\varphi_j^2\colon \mathbb{R}\times\Delta^2\to\mathscr{X}_{s,0}, j=1,2$, we have

$$\varphi_j^2(x,\tau_1,\tau_2)(0)=x\,,\quad \varphi_j^2(x,\tau_1,\tau_2)(s)=x-(-1)^j(s-2\tau_2)\,,$$

$$\varphi_j^2(x,\tau_1,\tau_2)(s_1)=\begin{pmatrix}x-(-1)^js_1 & (s_1<\tau_1)\,,\\x-(-1)^j(2\tau_1-s_1) & (\tau_1<s_1<\tau_1+\tau_2)\,,\\x-(-1)^j(s_1-2\tau_2) & (\tau_1+\tau_2<s_1)\,,\end{pmatrix}$$

so that we can get after all

$$\begin{aligned}\int_{\mathscr{X}_{s,0}} d\nu^2_{s,f;0,g}(X)\Psi(X)&=\int_{\mathbf{R}}dx\int_{\Delta^2}d\tau_1\,d\tau_2\sum_{j=1}^2(\overline{f(x-(-1)^j(s-2\tau_2))}N^{-1}B^2)_j(Ng)_j\Psi(\varphi_j^2(x,\tau_1,\tau_2))\\&=\sum_{j=1}^2\int_{\mathbf{R}\times\Delta^2}d\mu^{2,j}_{s,f;0,g}(x,\tau_1,\tau_2)\Psi(\varphi_j^2(x,\tau_1,\tau_2))\,.\end{aligned}$$

This proves Proposition 7 for $k=2$.

ACKNOWLEDGMENTS

The authors would like to thank Professor S. Kotani, Professor I. Mitoma, and Professor M. Tsuchiya for useful discussions.

This research was partially supported by Grants-in-Aid for Scientific Research, No. 61540091 and No. 62740118, from the Ministry of Education, Japan.

[1] T. Ichinose, Duke Math. J. **51**, 1 (1984); Proc. Jpn. Acad. A **58**, 290 (1982).

[2] T. Ichinose, Physica A **124**, 419 (1984).

[3] T. Ichinose and H. Tamura, J. Math. Phys. **25**, 1810 (1984).

[4] T. Ichinose and H. Tamura, "A remark on path integral for the Dirac equation," in *Proceedings of the Summer School on Functional Integration*, Sherbrooke, Canada, 1986 [a special issue of the Rend. Circ. Mat. Palermo (to appear)].

[5] R. P. Feynman and A. P. Hibbs, *Quantum Mechanics and Path Integrals* (McGraw-Hill, New York, 1965).

[6] G. V. Riazanov, Zh. Eksp. Teor. Fiz. **33**, 1437 (1958) [Soviet Phys. JETP **6**, 1107 (1958)].

[7] G. Rosen, *Formulations of Classical and Quantum Dynamical Theory* (Academic, New York, 1969), Appendix E, p. 118.

[8] E. Schrödinger, Sitz. Preuss. Akad. Wiss. Phys. Math. Kl. **24**, 418 (1930).

[9] Ph. Blanchard, Ph. Combe, M. Sirugue, and M. Sirugue-Collin, "Probabilistic solution of the Dirac equation; Path integral representation for the solution of the Dirac equation in presence of an electromagnetic field," BiBoS, Universität Bielefeld, preprint 1985.

[10] B. Gaveau, T. Jacobson, M. Kac, and L. S. Schulman, Phys. Rev. Lett. **53**, 419 (1984).

[11] B. Gaveau, J. Func. Anal. **58**, 310 (1984).

[12] T. Jacobson, J. Phys. A: Math. Gen. **17**, 2433 (1984).

[13] G. F. De Angelis, G. Jona-Lasinio, M. Serva, and N. Zanghi, J. Phys. A: Math. Gen. **19**, 865 (1986).

[14] See, e.g., M. Reed and B. Simon, *Methods of Modern Mathematical Physics, I: Functional Analysis* (Academic, New York, 1980), revised and enlarged ed.

[15] K. Swong, Math. Ann. **155**, 270 (1964).

[16] L. Schwartz, *Radon Measures on Arbitrary Topological Spaces and Cylindrical Measures* (Oxford U.P., London, 1973).

A supersymmetric generalization of von Neumann's theorem

H. Grosse
Institut für Theoretische Physik, Universität Wien, Wien, Austria
L. Pittner
Institut für Theoretische Physik, Universität Graz, Graz, Austria

(Received 1 July 1987; accepted for publication 9 September 1987)

A formulation of supersymmetric quantum mechanics is given and superunitary and generalized canonical transformations are defined acting in a module. Next it is assumed that there are operators that give an irreducible representation of the canonical commutation and anticommutation relations, respectively, and it is proved that two such representations are connected by a uniquely determined superunitary transformation, under suitable domain assumptions. This extends the well-known uniqueness theorem of von Neumann to canonical (anti-) commutation relations using anticommuting parameters.

I. INTRODUCTION

Although supersymmetry had been invented more than 13 years ago,[1] the study of systems with finitely many degrees of freedom, which have this symmetry, began only after 1981.[2-4] Moreover, supersymmetric quantum mechanics (SSQM) has been treated at the beginning only at a formal level. After the first attempt to introduce a rigorous framework in Ref. 5, a formulation of the axioms in terms of sesquilinear forms together with a discussion of classes of models as well as a description of the space in which superunitary transformations act, has been given in Ref. 6.

Here we extend these first steps in several directions. In Sec. II we review the description of the Hilbert space for f fermionic and f bosonic degrees of freedom. A Klein–Jordan–Wigner transformation yields the connection to a space being the tensor product of some separable Hilbert space times a Grassmann algebra.[6] In that space representations of Lie superalgebras yield quantum mechanical models.

The formulation of superunitary transformations needs more. The separable Hilbert space is generalized to a module that is obtained by using the Grassmann algebra of one Θ variable, over the field of complex numbers, as coefficients (Sec. III). Rules for evaluating sesquilinear forms are best formulated with the help of the Klein operator. An analog of Stones' theorem for superunitary groups is formulated, too.

We next discuss one-parameter groups of superunitary transformations in Sec. IV, which mix bosonic and fermionic operators but leave invariant the canonical (anti-) commutation relations [C(A)CR].

A natural question arises, which is at the origin of our analysis. Irreducible representations of the CCR are according to von Neumann's theorem unitarily equivalent to each other, and the same holds for the CAR.[7] Do similar results hold for irreducible operator representations of Lie superalgebras? Beside examples and corollaries our main result of Sec. V is formulated in Theorem 3. We start from operators fulfilling the C(A)CR and assume certain domain properties. We prove that two representations of the C(A)CR are connected by a unique superunitary transformation. Although the proof is straightforward it is somewhat tedious and put in Appendix A.

Following the same strategy we formulate in Sec. VI the extension of the above theorem to the case where one enlarges the original Hilbert space by the Grassmann algebra formed out of two anticommuting Θ variables; the proof is given in Appendix B.

II. THE HILBERT SPACE FOR f FERMIONIC AND f BOSONIC DEGREES OF FREEDOM

The Z_2 grading of the Hilbert space of states $\mathscr{H}_f$, for f fermionic and f bosonic degrees of freedom, may be obtained from the Grassmann algebra G_f of polynomials in the anticommuting variables $\varepsilon_1,...,\varepsilon_f$ over the field of complex numbers $\mathbb{C}$. Any element $\xi \in G_f$ can be expanded into the 2^f monomials[8]

$$\xi = c_0 I + \sum_{1<i_1<\cdots<i_p<f} c_{i_1\cdots i_p}\varepsilon_{i_1}\cdots\varepsilon_{i_p}, \quad c_0, c_{i_1\cdots i_p} \in \mathbb{C}, \tag{2.1}$$

where I denotes the unit element of the associative superalgebra G_f. Here ξ is called even (odd), if it is the linear combination of monomials $\varepsilon_{i_1}\cdots\varepsilon_{i_p}$ with even (odd) p.

The derivative from the left $\partial_k = \partial/\partial\varepsilon_k$ with respect to ε_k is defined by the linear extension of

$$\begin{aligned} &\partial_k \varepsilon_{i_1}\cdots\varepsilon_{i_p} = \sum_{l=1}^{p} \delta_{kl}(-)^{l-1}\varepsilon_{i_1}\cdots\not{\varepsilon}_{i_l}\cdots\varepsilon_{i_p}, \\ &\partial_k I = 0, \quad k = 1,...,f, \end{aligned} \tag{2.2}$$

where $\not{\varepsilon}_m$ means that ε_m has to be omitted. For homogeneous elements $\xi \in \mathscr{G}_f$, the degree $\deg \xi$ is defined to be zero (one) if ξ is even (odd); one therefore obtains the rule that

$$\partial_k(\xi n) = (\partial_k \xi)\eta + (-)^{\deg \xi}\xi(\partial_k \eta), \quad k = 1,...,f. \tag{2.3}$$

The Grassmann algebra $\mathscr{G}'_f$ of polynomials in $\partial_1,...,\partial_f$ may be combined with $\mathscr{G}_f$ to yield the Clifford algebra K_{2f} of polynomials over $\mathbb{C}$ in the variables ε_k and ∂_k, $k = 1,...,f$, which fulfill the canonical anticommutation relations (CAR)

$$\{\varepsilon_i, \varepsilon_k\} = \{\partial_i, \partial_k\} = 0, \quad \{\varepsilon_i, \partial_k\} = \delta_{ik} I, \quad i,k = 1,...,f. \tag{2.4}$$

0022-2488/88/010110-09$02.50

The algebra (2.4) can be represented on the f-fold tensor product $\overset{f}{\otimes}\mathbb{C}^2$ by Pauli matrices $\sigma^{\pm},\sigma^3$,

$$C_k = (-\sigma^3)\otimes\cdots\otimes(-\sigma^3)\otimes\underset{\substack{\uparrow\\ k\text{ th place}}}{\sigma^-}\otimes I_2\otimes\cdots\otimes I_2,\quad k=2,\ldots \tag{2.5}$$

$$C_1=\sigma^-\otimes I_2\otimes\cdots\otimes I_2,\quad I_2=\begin{pmatrix}1&0\\0&1\end{pmatrix},\quad \sigma^-=\begin{pmatrix}0&0\\1&0\end{pmatrix},$$

and the isomorphism $C_k^\dagger\leftrightarrow\varepsilon_k$ and $C_k\leftrightarrow\partial_k$, $k=1,\ldots,f$, holds. With the scalar product[6]

$$\langle\xi\,|\eta\rangle=c_0^*d_0+\sum_{1<i_1<\cdots<i_p\leqslant f}c^*_{i_1\cdots i_p}d_{i_1\cdots i_p}, \tag{2.6}$$

the Grassmann algebra $\mathscr{G}_f$ becomes isomorphic to the unitary space $\mathbb{C}^{2^f}$, and $\partial_k=\varepsilon_k^\dagger$, $k=1,\ldots,f$.

The Hilbert space of SSQM is now defined by the tensor product $\mathscr{H}_f=\mathscr{H}_0\otimes\mathscr{G}_f$, where $\mathscr{H}_0$ denotes some separable Hilbert space. From the Z_2 grading of $\mathscr{G}_f$ one obtains an orthogonal decomposition $\mathscr{H}_f=\mathscr{H}_f^0\oplus\mathscr{H}_f^1$ into even and odd elements, with projection operators N_0 and N_1, projecting onto bosonic and fermionic states; $K=N_0-N_1=(-1)^{N_1}$ denotes the Klein operator.[9]

The scalar product for states $\Psi,\Phi\in\mathscr{H}_f$ with

$$\Psi=\psi_0I+\sum_{1<i_1<\cdots<i_p\leqslant f}\varepsilon_{i_1}\cdots\varepsilon_{i_p}\psi_{i_1\cdots i_p},\quad \psi_{i_1\cdots i_p},\psi_0\in\mathscr{H}_0, \tag{2.7}$$

and Φ similarly, is defined by

$$\langle\Phi|\Psi\rangle=\langle\phi_0|\psi_0\rangle+\sum_{1<i_1<\cdots<i_p\leqslant f}\langle\phi_{i_1\cdots i_p}|\psi_{i_1\cdots i_p}\rangle. \tag{2.8}$$

For f bosonic degrees of freedom we may take $\mathscr{H}_0=L^2(d^fx)$ and denote $\mathscr{L}_f=L^2(d^fx)\otimes\mathscr{G}_f$. The obvious isomorphism $L_f\leftrightarrow\overset{f}{\otimes}(L^2(d^1x)\otimes\mathbb{C}^2)$ shows that f distinguishable fermions on the real line are described by Pauli spinors.

In $L^2(d^fx)$, the closed operators

$$B_k=\frac{x_k+\partial/\partial x_k}{\sqrt2},\quad B_k^\dagger=\frac{x_k-\partial/\partial x_k}{\sqrt2},\quad k=1,\ldots,f,$$
$$\operatorname{dom}B_k=\operatorname{dom}B_k^\dagger=\operatorname{dom}x_k\cap\operatorname{dom}p_k, \tag{2.9}$$
$$p_k=-i\frac{\partial}{\partial x_k},$$

fulfill the canonical commutation relations (CCR)

$$[B_i,B_k]=0,\quad [B_i,B_k^\dagger]=\delta_{ik}I,\quad i,k=1,\ldots,f, \tag{2.10}$$

in the sense of sesquilinear forms.[10] Taking the tensor product of them with the bounded operators ε_k and ∂_k on $\mathscr{G}_f$ yields self-adjoint supercharges

$$Q^1=\frac{Q+Q^\dagger}{\sqrt2},\quad Q^2=\frac{-i(Q-Q^\dagger)}{\sqrt2},$$
$$Q=\sqrt2\sum_{k=1}^{f}B_k\varepsilon_k, \tag{2.11}$$

and the non-negative Hamilton operator of f fermionic oscillators

$$H=\sum_{k=1}^{f}H_k,$$
$$H_k=B_k^\dagger B_k+\varepsilon_k\,\partial_k=\tfrac12(p_k^2\dotplus x_k^2+\sigma_k^3)\geqslant0, \tag{2.12}$$
$$\operatorname{dom}H^{1/2}=\operatorname{dom}Q^1=\operatorname{dom}Q^2=\bigcap_{k=1}^{f}\operatorname{dom}B_k\otimes G_f;$$

here the representation (2.5) is used to represent the spin-flip energy

$$\tfrac12\sigma_k^3=\tfrac12I_2\otimes\cdots\otimes I_2\otimes\sigma^3\otimes I_2\cdots\otimes I_2.$$

One therefore obtains, in the sense of forms, the following representation of the Lie superalgebra $S(2)$ by self-adjoint operators in $\mathscr{L}_f$:

$$(Q^1)^2=(Q^2)^2=H,\quad \{Q^1,Q^2\}=0,$$
$$\{Q^1,K\}=\{Q^2,K\}=0 \tag{2.13}$$

or, equivalently,

$$Q^2=0,\quad \{Q,Q^\dagger\}=2H,\quad \{Q,K\}=0. \tag{2.14}$$

Every irreducible representation of the CCR for f bosonic degrees of freedom in a separable Hilbert space is unitarily equivalent to the harmonic oscillator representation (2.10) according to von Neumann's theorem[7]; similarly every irreducible representation of the CAR for f fermionic degrees of freedom in a separable Hilbert space is unitarily equivalent to the representation (2.5). These two uniqueness theorems can be generalized to superunitary transformations that mix bosonic and fermionic operators, using anticommuting parameters.

III. SUPERUNITARY TRANSFORMATIONS

Whereas at least two anticommuting parameters are necessary in order to construct the Lie supergroup corresponding to the Lie superalgebra (2.13),[11] one needs only one such parameter Θ for an appropriate definition of superunitary transformations of CCR and CAR. The skew-symmetric tensor product of the Clifford algebra K_{2f} with the Grassmann algebra $\mathscr{D}_1$ of polynomials in Θ with complex coefficients[12] is an associative superalgebra over $\mathbb{C}$, and also a Lie superalgebra with anticommutation rules $\{\Theta,\varepsilon_k\}=\{\Theta,\partial_k\}=0$, $k=1,\ldots,f$, and $\Theta^2=0$.

The Hilbert space $\mathscr{H}_f$ is extended to the $\mathscr{D}_1$ module $\mathscr{H}_f\oplus\Theta\mathscr{H}_f=\mathscr{H}_f(\Theta)$ (Ref. 13) with elements $\Psi=\Psi_0+\Theta\Psi_1$, Ψ_0 and $\Psi_1\in\mathscr{H}_f$. The scalar product on $\mathscr{H}_f\times\mathscr{H}_f$ is generalized to a sesquilinear mapping from $\mathscr{H}_f(\Theta)\times\mathscr{H}_f(\Theta)$ onto $\mathscr{D}_1$[9]:

$$\langle\Theta\Phi|\Psi\rangle=\langle\Phi|\Theta\Psi\rangle=\Theta\langle\Phi|K\Psi\rangle,\quad \Phi,\Psi\in\mathscr{H}_f,$$
$$(\Theta\varepsilon_k)^\dagger=\partial_k\Theta. \tag{3.1}$$

A densely defined linear operator A in $\mathscr{H}_f$ may be decomposed into an even and odd part,

$$A=A_0+A_1,\quad A_0=N_0AN_0+N_1AN_1,$$
$$A_1=N_0AN_1+N_1AN_0, \tag{3.2}$$

with $\operatorname{dom}A_0=\operatorname{dom}A_1=\operatorname{dom}A$, under an additional assumption: for all $\Psi \in \operatorname{dom} A$, $K\Psi \in \operatorname{dom} A$. The domain of A is then called graded (for the notion of a graded subspace see Ref. 12). The adjoint operator can then be decomposed into

$$\begin{aligned} &A^\dagger = A_0^\dagger + A_1^\dagger, \quad A_0^\dagger = N_0 A^\dagger N_0 + N_1 A^\dagger N_1, \\ &A_1^\dagger = N_0 A^\dagger N_1 + N_1 A^\dagger N_0, \end{aligned} \tag{3.3}$$

with $\operatorname{dom} A_0^\dagger = \operatorname{dom} A_1^\dagger = \operatorname{dom} A^\dagger$, if $\operatorname{dom} A^\dagger$ is graded, too; this we shall assume from now on. With the definitions

$$A\Theta = \Theta KAK = \Theta(A_0 - A_1), \quad AK = K(A_0 - A_1),$$

and (3.4)

$$(\Theta A)^\dagger = \Theta K A^\dagger K = \Theta(A_0^\dagger - A_1^\dagger) = A^\dagger \Theta, \quad \Theta K = K\Theta,$$

one obtains immediately the rules

$$\begin{aligned} \langle \Phi | \Theta A \Psi \rangle &= \theta \langle \Phi | K A \Psi \rangle \\ &= \Theta \langle \Phi | (A_0 - A_1) K \Psi \rangle \\ &= \Theta \langle (A_0^\dagger - A_1^\dagger) \Phi | K \Psi \rangle \\ &= \Theta \langle A^\dagger K \Phi | \Psi \rangle = \langle (A_0^\dagger - A_1^\dagger) \Phi | \Theta \Psi \rangle \\ &= \langle A^\dagger \Theta \Phi | \Psi \rangle, \quad \Psi \in \operatorname{dom} A, \quad \Phi \in \operatorname{dom} A^\dagger. \end{aligned} \tag{3.5}$$

In order to construct superunitary transformations, we consider the operator family

$$\begin{aligned} &g(t) = I + \Theta t A, \quad A = A_0 + A_1, \quad A_0 = iA_0', \\ &g^\dagger(t) = I + tA^\dagger \Theta = I - it\Theta A_0'^\dagger - t\Theta A_1^\dagger, \quad t \text{ real.} \end{aligned} \tag{3.6}$$

If A_0' and A_1 are symmetric, then obviously the products

$$g(t) g^\dagger(t) = g^\dagger(t) g(t) = I|_{\operatorname{dom} A}, \quad t \text{ real.} \tag{3.7}$$

This transformation $g(t)$ may be extended to a domain $\operatorname{dom} A \oplus \Theta \mathscr{H}_f$ of states $\Psi_0 + \Theta \Psi_1$ with $\Psi_0 \in \operatorname{dom} A$ and $\Psi_1 \in \mathscr{H}_f$, which is usually written as $g(t) = \exp(t\Theta A)$, t real. Operators $B + \Theta A$ are defined on the domain $\operatorname{dom}(B + \Theta A) = \operatorname{dom}(B + A) \oplus \Theta \operatorname{dom} B$, which is a subspace of the above-defined $\mathscr{D}_1$ module; $\operatorname{dom} B$ and $\operatorname{dom} A$ are assumed here to be graded; then $(B + \Theta A)^\dagger = B^\dagger + A^\dagger \Theta$, if $\operatorname{dom} B^\dagger$ and $\operatorname{dom} A^\dagger$ are graded, too.

Definition 1: The transformation $g(t) = I + \Theta t A$ with t real is called superunitary iff A is odd ($A_0 = 0$) and self-adjoint ($A_1 = A_1^\dagger$); then $\operatorname{dom} g(t) = \operatorname{dom} g^\dagger(t) = \operatorname{dom} A \oplus \Theta \mathscr{H}_f$.

The operator family (3.6) fulfills the group multiplication law $g(t) g(s) = g(t+s)$, t and s real. Conversely, the following analog of Stones' theorem holds.

Lemma 1: Consider an operator group $g(t) = I + \Theta A(t)$ with $A(t)$ densely defined in $\mathscr{H}_f$, which fulfills the group property $g(t) g(s) = g(t+s)$, t and s real, $g(0) = I$, in the sense of operator sums $A(t+s) = A(t) + A(s)$, $A(0) = 0$. Assume that $\operatorname{dom} \mathrm{A}(t) = \operatorname{dom} A(-t)$, for $t > 0$; then $\operatorname{dom} A(t) = \operatorname{dom} A$, $A = A(1)$, for all t. If the mapping $\{\langle \Phi | A(t) \Psi \rangle; t \in \mathbb{R}\}$ is continuous for all $\Phi, \Psi \in \operatorname{dom} A$, then $A(t) = tA$ holds for all $t \in \mathbb{R}$.

Lemma 1 follows since $A(t) - tA(1) = 0$ for all rational numbers t.

If the unitary operator U is decomposed according to (3.2), $U = U_0 + U_1$, then $KUK = U_0 - U_1$ is unitary, too; one obtains $U_0^\dagger U_0 + U_1^\dagger U_1 = U_0 U_0^\dagger + U_1 U_1^\dagger = I$.

IV. GENERALIZED CANONICAL TRANSFORMATIONS

The one-parameter group of superunitary transformations

$$\begin{aligned} &g(t) = I + \Theta t A, \\ &AK = -KA \quad \text{on } K \cdot \operatorname{dom} A = \operatorname{dom} A, \quad A = A^\dagger, \\ &g^\dagger(t) = I - \Theta t A, \quad g^\dagger(t) g(t) = g(t) g^\dagger(t) = I|_{\operatorname{dom} g(t)}, \\ &g(t+s) = g(t) g(s), \quad t \text{ and } s \text{ real}, \\ &\operatorname{dom} g(t) = \operatorname{dom} A \oplus \Theta \mathscr{H}_f, \end{aligned} \tag{4.1}$$

may be used to perform generalized canonical Bogoliubov transformations, which mix bosonic and fermionic operators:

$$\begin{aligned} &B_k'(t) = g(t) B_k g^\dagger(t) = B_k + \Theta t\, [A, B_k], \\ &\operatorname{dom} B_k'(t) = \operatorname{dom}\, [A, B_k] \oplus \Theta \operatorname{dom} B_k, \\ &C_k'(t) = g(t) \partial_k g^\dagger(t) = \partial_k + \Theta t \{A, \partial_k\}, \\ &\operatorname{dom} \mathrm{C}_k'(t) = \operatorname{dom} \{A, \partial_k\} \oplus \Theta \mathscr{H}_f, \end{aligned} \tag{4.2}$$

for t real and $k = 1, \ldots, f$. The adjoint operators are given by

$$\begin{aligned} B_k'^\dagger(t) &= B_k^\dagger - \Theta t\, [A, B_k]^\dagger \supseteq B_k^\dagger + \Theta t\, [A, B_k^\dagger] \\ &= g(t) B_k^\dagger g^\dagger(t), \\ C_k'^\dagger(t) &= \varepsilon_k + \Theta t \{A, \partial_k\}^\dagger \supseteq \varepsilon_k + \Theta t \{A, \varepsilon_k\} \\ &= g(t) \varepsilon_k g^\dagger(t). \end{aligned} \tag{4.3}$$

Note that $\operatorname{dom} A$ is assumed to be graded, which in turn implies that the domains of $[A, B_k]$, $[A, B_k^\dagger]$, $\{A, \partial_k\}$, and $\{A, \varepsilon_k\}$ are graded, too.

These transformed operators, acting in the $\mathscr{D}_1$ module $\mathscr{H}_f \oplus \Theta \mathscr{H}_f$, obey again the CCR and CAR in the sense of sesquilinear forms, with domains of $B_k'^\dagger(t)$ and $C_k'^\dagger(t)$ restricted, according to (4.3), to $\operatorname{dom}\, [A, B_k^\dagger] \oplus \Theta \operatorname{dom} B_k$ and $\operatorname{dom} \{A, \varepsilon_k\} \oplus \Theta \mathscr{H}_f$, respectively:

$$\begin{aligned} &\langle B_i'^\dagger(t) \cdot | B_k'^\dagger(t) \cdot \rangle - \langle B_k'(t) \cdot | B_i'(t) \cdot \rangle = \delta_{ik} \langle \cdot | \cdot \rangle, \\ &\langle C_i'^\dagger(t) \cdot | C_k'^\dagger(t) \cdot \rangle + \langle C_k'(t) \cdot | C_i'(t) \cdot \rangle = \delta_{ik} \langle \cdot | \cdot \rangle, \end{aligned} \tag{4.4}$$

and similar (anti-) commutation relations for $C_i'(t)$ with $C_k'(t)$ and $B_i'(t)$ with $B_k'(t)$ and between them. These relations cannot be extended to the domains of $B_k'^\dagger(t)$ and $C_k'^\dagger(t)$, in general.

The form invariance of the C(A)CR under superunitary transformations is at the origin of our attempts to extend von Neumann's uniqueness theorem for irreducible representations of the CCR and the corresponding theorem for the CAR[7] to representations of the C(A)CR in the $\mathscr{D}_1$ module $\mathscr{H}_f \oplus \Theta \mathscr{H}_f$.

V. UNIQUENESS OF C(A)CR IN $\mathscr{H}_f \oplus \Theta \mathscr{H}_f$

An irreducible representation of the canonical (anti-) commutation relations by closed operators in $\mathscr{H}_f$ is, up to unitary and superunitary transformations, unique, in the following sense:

Theorem 1: Let C_k, $k = 1,\ldots,f$, be bounded operators on the separable Hilbert space $\mathscr{H}$, and let B_k, $k = 1,\ldots,f$, be densely defined closed operators in $\mathscr{H}$. Assume that the canonical (anti-) commutation relations [C(A)CR] hold:

$$C_i C_k^\dagger + C_k^\dagger C_i = \delta_{ik} I, \quad C_i C_k + C_k C_i = 0, \quad i,k = 1,\ldots,f,$$
$$B_k B_k^\dagger - B_k^\dagger B_k \subseteq I, \quad \operatorname{dom} B_k B_k^\dagger = \operatorname{dom} B_k^\dagger B_k. \tag{5.1a}$$

On $\operatorname{dom} B_k = \operatorname{dom} B_k^\dagger, (B_k + B_k^\dagger)/\sqrt{2}$ and $i(B_k^\dagger - B_k)/\sqrt{2}$ are essentially self-adjoint; their closures will be denoted by X_k and P_k. Assume next that the commutation relations

$$\begin{gathered} B_i B_k - B_k B_i = 0, \quad B_i B_k^\dagger - B_k^\dagger B_i = 0, \\ i \neq k, \quad i,k = 1,\ldots,f, \end{gathered} \tag{5.1b}$$

hold for the spectral families of X_i, P_i. In addition

$$B_i C_k - C_k B_i = 0, \quad B_i^\dagger C_k - C_k B_i^\dagger = 0, \quad i,k = 1,\ldots,f, \tag{5.1c}$$

should hold for these spectral families. Finally, the operator family $\{B_k, B_k^\dagger, C_k, C_k^\dagger; k = 1,\ldots,f\}$ is assumed to be irreducible in $\mathscr{H}$, which means that there does not exist a nontrivial invariant closed linear subspace of $\mathscr{H}$ for this family.

Then there is a unitary transformation U: $\mathscr{H} \leftrightarrow \overset{f}{\otimes} (L^2(d^1x) \otimes \mathbb{C}^2)$ such that

$$\begin{aligned} &U B_k U^{-1} = (x_k + i p_k)/\sqrt{2}, \quad k = 1,\ldots,f, \\ &U C_k U^{-1} = (-\sigma^3) \otimes \cdots \otimes (-\sigma^3) \otimes \underset{\overset{\uparrow}{k\text{th place}}}{\sigma^-} \otimes I_2 \otimes \cdots \otimes I_2, \\ &\qquad \text{for } k = 2,\ldots,f, \\ &U C_1 U^{-1} = \sigma^- \otimes I_2 \otimes \cdots \otimes I_2. \end{aligned} \tag{5.2}$$

Remark: It follows that each representation of the C(A)CR in the sense of (5.1) is, up to unitary equivalence, the direct sum of countable many fermionic oscillator families (5.2).[7] This uniqueness theorem can be extended to the following representations of C(A)CR in the $\mathscr{D}_1$ module $\mathscr{H}_f \oplus \Theta \mathscr{H}_f$. We start with the case of one fermion, $f = 1$.

Theorem 2: Let the linear operators B and C fulfill all conditions of Theorem 1 with $f = 1$, such that one may identify $B = (x + ip)/\sqrt{2}$ and $C^\dagger = \varepsilon$, $C = \partial = \partial/\partial\varepsilon$ in $\mathscr{H} = \mathscr{H}_1 = \mathscr{H}_0 \otimes \mathscr{G}_1 = L^2(d^1x) \otimes \mathscr{G}_1$. Let D and G be densely defined closed linear operators. Define

$$\begin{aligned} &B' = B + \Theta D, \\ &\operatorname{dom} B' = ((\operatorname{dom} B \otimes \mathscr{G}_1) \cap \operatorname{dom} D) \\ &\qquad \oplus \Theta(\operatorname{dom} B \otimes \mathscr{G}_1), \\ &C' = \partial + \Theta G, \quad \operatorname{dom} C' = \operatorname{dom} G \oplus \Theta \mathscr{H}_1, \\ &B'^\dagger = B^\dagger - \Theta D^\dagger, \quad C'^\dagger = \varepsilon + \Theta G^\dagger, \end{aligned} \tag{5.3}$$

where $G(D)$ is assumed to be even (odd); we may identify $\varepsilon = \sigma^\dagger, \partial = \sigma^-$ on $\mathscr{G}_1 = \mathbb{C}^2$. The new operators are assumed to fulfill the C(A)CR in the sense of operator polynomials, i.e.,

$$C'^2 = 0, \quad [B', C'] = 0, \quad [B', C'^\dagger] = 0. \tag{5.4}$$

Moreover, the following domain conditions are assumed. With

$$G = G_{00} I_2 + G_{11} \varepsilon \partial, \quad D = D_{10} \varepsilon + D_{01} \partial, \tag{5.5}$$

such that D_{10}, D_{01}, G_{00}, and $G_{00} + G_{11}$ are closed and densely defined in $\mathscr{H}_0$, assume that

$$\operatorname{dom} G_{00} = \operatorname{dom} G_{00}^\dagger, \quad \operatorname{dom} G_{00} \subseteq \operatorname{dom} G_{11}, \tag{5.6}$$

$$\begin{aligned} &\operatorname{dom} D_{10} \cap \operatorname{dom} D_{01} \text{ is a core of both } D_{10} \text{ and } D_{01}, \\ &\operatorname{dom} [B, G_{00}] \cap \operatorname{dom} D_{10} \cap \operatorname{dom} D_{01} \\ &\qquad \text{is a core of both } \overline{[B, G_{00}]} \text{ and } D_{10}, \\ &\operatorname{dom} [B, G_{00}^\dagger] \operatorname{dom} D_{10} \cap \operatorname{dom} D_{01} \\ &\qquad \text{is a core of both } \overline{[B, G_{00}^\dagger]} \text{ and } D_{01}, \end{aligned} \tag{5.7}$$

where $\overline{T}$ denotes the closure of the operator T. Then one can write

$$\begin{aligned} &G = G_{00} I = \{A, \partial\}, \\ &D = \overline{[A, B]} \quad \text{with } A = G_{00} \varepsilon + G_{00}^\dagger \partial = A^\dagger, \end{aligned} \tag{5.8}$$

and the transformation (5.3) is implemented by the superunitary operator $\exp(\Theta A)$ such that

$$\begin{aligned} &e^{\Theta A} B e^{-\Theta A} = B + \Theta[A, B] \subseteq B', \\ &C' = e^{\Theta A} \partial e^{-\Theta A} = \partial + \Theta\{A, \partial\}. \end{aligned} \tag{5.9}$$

Corollary 1: Under the conditions of the above theorem,

$$D_{10} = \overline{[G_{00}, B]} \quad \text{and} \quad D_{01} = \overline{[G_{00}^\dagger, B]}.$$

Remark: Tensor products of operators $L \otimes I$ with $\operatorname{dom} L \subseteq \mathscr{H}_0$ will be written sometimes as L, and the identity mapping I of $\mathscr{G}_f$ will be suppressed.

Corollary 2: There exists only one self-adjoint odd operator that generates the superunitary transformation (5.9), namely A defined above.

Proof of Theorem 2: With the ansatz (5.5), $C'^2 = 0$ and (5.6) implies that $G = G_{00} I_2$. Here A is self-adjoint because G_{00} is closed; (5.6) implies that $G = \{A, \partial\}$; $[B', C'] = 0$, $[B', C'^\dagger] = 0$, and (5.7) leads to $D = \overline{[A, B]}$.

Corollary 3: Under the conditions of the above theorem,

$$\overline{[D, B^\dagger]} \supseteq [[G, B], B^\dagger]\varepsilon + [[G^\dagger, B], B^\dagger]\partial \subseteq [B, D^\dagger] \tag{5.10}$$

holds; moreover, the equivalence

$$[B', B'^\dagger] \subseteq I \quad \text{iff} \quad [D, B^\dagger] - [B, D^\dagger] = 0 \tag{5.11}$$

follows. The conditions

$$\begin{aligned} &\operatorname{dom} [[G, B], B^\dagger] \\ &\qquad \text{is a core for both } \overline{[\overline{[G, B]}, B^\dagger]} \text{ and } \overline{[B, [G^\dagger, B]^\dagger]}, \\ &\operatorname{dom} [[G^\dagger, B], B^\dagger] \\ &\qquad \text{is a core for both } [\overline{[G^\dagger, B]}, B^\dagger] \text{ and } [B, [G, B]^\dagger], \end{aligned} \tag{5.12}$$

imply that $\overline{[D, B^\dagger]} = \overline{[B, D^\dagger]}$ and *a fortiori* $[B', B'^\dagger] \subseteq I$.

Example 1: If one chooses $G = cB$ with c complex, $D = -c^*\partial$ and one obtains the superunitary transformation[6]

$$\begin{aligned} &\begin{pmatrix} B' \\ C' \end{pmatrix} = \begin{pmatrix} I & -c^*\Theta \\ c\Theta & I \end{pmatrix} \begin{pmatrix} B \\ \partial \end{pmatrix}, \\ &\begin{pmatrix} I & -c^*\Theta \\ c\Theta & I \end{pmatrix} \begin{pmatrix} I & c^*\Theta \\ -c\Theta & I \end{pmatrix} = \begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix}. \end{aligned} \tag{5.13}$$

Example 2: For $G = cB^\dagger$, c complex, $D = -c\varepsilon$.

An extension of the above theorem to the case of f fermionic and f bosonic degrees of freedom follows from a similar strategy as before. We require the C(A)CR as operator polynomials and assume appropriate domain conditions.

Theorem 3: Let C_k be bounded, B_k, G_k, and D_k be closed and densely defined operators in $\mathscr{H}$, for $k = 1,\ldots,f$. Decompose $\mathscr{H}$ into two infinite-dimensional orthogonal subspaces $\mathscr{H} = \mathscr{H}^0 \oplus \mathscr{H}^1$ and let B_k and G_k be even, C_k and D_k be odd. Define

$$B'_k = B_k + \Theta D_k, \quad C'_k = C_k + \Theta G_k, \quad k = 1,\ldots,f, \tag{5.14}$$

and note that

$$B'^{\dagger}_k = B^\dagger_k - \Theta D^\dagger_k, \quad C'^{\dagger}_k = C^\dagger_k + \Theta G^\dagger_k \tag{5.15}$$

follows, since the domains of the operators involved are graded. Assume

$$\begin{gathered}[B'_i, B'^{\dagger}_k] \subseteq \delta_{ik} I, \quad [B'_i, B'_k] = 0, \quad [B'_i, C'_k] = 0,\\ \{C'_i, C'^{\dagger}_k\} \subseteq \delta_{ik} I, \quad \{C'_i, C'_k\} = 0, \quad [B'_i, C'^{\dagger}_k] = 0,\\ i,k = 1,\ldots,f.\end{gathered} \tag{5.16}$$

Let $\operatorname{dom} B_k B^\dagger_k = \operatorname{dom} B^\dagger_k B_k$ and define X_k, P_k as in Theorem 1. Assume that the commutation relations

$$\begin{gathered}[B_i, B_k] = 0 \quad \text{and} \quad [B_i, B^\dagger_k] = 0, \quad \text{for } i \neq k,\\ [B_i, C_k] = 0 \quad \text{and} \quad [B_i, C^\dagger_k] = 0, \quad \text{for } i,k = 1,\ldots,f,\end{gathered} \tag{5.17}$$

which hold according to (5.16) in the sense of operator polynomials, hold also in the sense of the spectral resolutions of X_i, P_i. Assume in addition that the family $\{B_k, B^\dagger_k, C_k, C^\dagger_k;\ k = 1,\ldots,f\}$ acts irreducibly in $\mathscr{H}$, such that we may identify, according to Theorem 1,

$$\begin{gathered}\mathscr{H} \leftrightarrow L_f, \quad B_k \leftrightarrow (x_k + ip_k)/\sqrt{2}, \quad C^\dagger_k \leftrightarrow \varepsilon_k,\\ C_k \leftrightarrow \partial_k, \quad k = 1,\ldots,f,\end{gathered} \tag{5.18}$$

by an appropriate unitary transformation. Introduce components for G_k by writing

$$\begin{aligned}G_k = G^0_k &+ \sum_{1<p_1<\cdots<p_t<f} G^{p_1\cdots p_t}_k \varepsilon_{p_1}\cdots\varepsilon_{p_t}\\ &+ \sum_{1<q_1<\cdots<q_u<f} G^{q_1\cdots q_u}_k \partial_{q_1}\cdots\partial_{q_u}\\ &+ \sum_{\substack{1<p_1<\cdots<p_r<f\\ 1<q_1<\cdots<q_s<f}} G^{p_1\cdots p_r, q_1\cdots q_s}_k \varepsilon_{p_1}\cdots\varepsilon_{p_r}\partial_{q_1}\cdots\partial_{q_s},\\ &t,u,r+s \text{ even},\end{aligned} \tag{5.19}$$

and assume that there is some domain

$$\begin{gathered}\mathscr{C} \subseteq \bigcap_{\substack{\{p,q\}\\ k=1}}^{f} (\operatorname{dom} G^{\{p,q\}}_k \cap \operatorname{dom} G^{\{p,q\}\dagger}_k),\\ \mathscr{C} \otimes \mathscr{G}_f \subseteq \operatorname{dom} D_k \cap \operatorname{dom} D^\dagger_k,\end{gathered} \tag{5.20}$$

such that $\mathscr{C}_f = \mathscr{C} \otimes \mathscr{G}_f$ is a core for G_k and D_k, $k = 1,\ldots,f$. Here $\{p,q\}$ denotes the 2^{2f} combinations in (5.19). Moreover, assume

$$\begin{gathered}\mathscr{C} \subseteq \operatorname{dom} B_k, \quad B_k\mathscr{C} \cup B^\dagger_k\mathscr{C} \subseteq \mathscr{C},\\ G^{\{p,q\}}_k\mathscr{C} \cup G^{\{p,q\}\dagger}_k\mathscr{C} \subseteq \operatorname{dom} B_k.\end{gathered} \tag{5.21}$$

Then the following conclusion holds. There exists exactly one densely defined operator $\dot{A}$ such that $\dot{A}$ is symmetric, $\operatorname{dom} \dot{A} = \mathscr{C}_f$,

$$\begin{gathered}\{\dot{A}, C_k\} = \dot{G}_k = G_{k|_{\mathscr{C}_f}}, \quad [\dot{A}, B_k] = \dot{D}_k = D_{k|_{\mathscr{C}_f}},\\ k = 1,\ldots,f.\end{gathered} \tag{5.22}$$

Then obviously

$$G_k = \overline{\{\dot{A}, C_k\}}, \quad D_k = \overline{[\dot{A}, B_k]}, \quad k = 1,\ldots,f. \tag{5.23}$$

The proof of this theorem is given in Appendix A.

Corollary 4: Under the conditions of the above theorem,

$$\begin{gathered}e^{\Theta\dot{A}} B_k e^{-\Theta\dot{A}} = B_k + \Theta\dot{D}_k \subseteq B'_k,\\ e^{\Theta\dot{A}} C_k e^{-\Theta\dot{A}} = C_k + \Theta\dot{G}_k \subseteq C'_k, \quad k = 1,\ldots,f.\end{gathered} \tag{5.24}$$

Remark: The transformation (5.24) need not be superunitary, because $\dot{A}$ may only be symmetric. The explicit representation of $\dot{A}$ in terms of components $G^{\{p,q\}}_k$, which is used in the subsequent proof, may be used to impose conditions such that $\dot{A}$ becomes essentially self-adjoint.

Remark: For practical use conditions (5.21) might be strengthened. One may require the existence of an invariant domain $\mathscr{C}_f$, for instance, $\mathscr{C} = S(\mathbb{R}^f)$, the rapidly decreasing C^∞ functions, for all the operators involved.

Example 3: The essentially self-adjoint operator

$$\dot{A} = \sum_{k=1}^{f} (\varepsilon_k \dot{B}_k + \partial_k B^\dagger_k), \quad \dot{B}_k = B_{k|_{\mathscr{C}}}, \quad \mathscr{C} = C^\infty_0(\mathbb{R}^f), \tag{5.25}$$

generates

$$\{\dot{A}, \partial_k\} = \dot{B}_k, \quad [\dot{A}, B_k] = -\partial_{k|_{C\otimes G_f}}, \quad k = 1,\ldots,f, \tag{5.26}$$

which leads to the superunitary matrix transformation[6]

$$B'_k = B_k - \Theta\partial_k, \quad C'_k = \partial_k + \Theta B_k, \quad k = 1,\ldots,f. \tag{5.27}$$

The closure of this generator $\dot{A}$ is just one of the two self-adjoint supercharges of the f-dimensional fermionic oscillator,

$$\begin{gathered}Q^1 = \overline{\dot{A}} = \sum_{k=1}^{f} (\varepsilon_k B_k + \partial_k B^\dagger_k),\\ \operatorname{dom} Q^1 = \operatorname{dom} B_k \otimes \mathscr{G}_f,\end{gathered} \tag{5.28}$$

and obviously

$$\{Q^1, \partial_k\} = B_{k|_{\bigcap_{l=1}^{f} \operatorname{dom} B_l \otimes G_f}}, \quad [Q^1, B_k] \subsetneqq -\partial_k. \tag{5.29}$$

Similar matrix transformations are generated by the essentially self-adjoint operators

$$\begin{gathered}\dot{A} = \sum_{k=1}^{f} (\varepsilon_k \dot{G}_k + \partial_k G^\dagger_k), \quad \dot{G}_k = -i\dot{B}_k \quad \text{or} \quad iB^\dagger_{k|_{\mathscr{C}}},\\ \overline{\dot{A}} = \sum_{k=1}^{f} (\varepsilon_k G_k + \partial_k G^\dagger_k), \quad G_k = \overline{\dot{G}_k},\\ \operatorname{dom} \overline{\dot{A}} = \operatorname{dom} B_k \otimes \mathscr{G}_f.\end{gathered} \tag{5.30}$$

VI. TRANSFORMATIONS WITH TWO ANTICOMMUTING PARAMETERS

More generally, superunitary transformations with more than one anticommuting parameter can be studied. They lead to Lie superalgebras that are much more complicated compared to the one of Eqs. (A3). As an example consider the supergroup constructed from the Lie superalgebra $S(2)$ with two self-adjoint supercharges Q^1 and Q^2 acting in the Hilbert space $\mathscr{H}_f$[6]:

$$(Q^1)^2 = (Q^2)^2 = H, \quad \operatorname{dom} Q^1 = \operatorname{dom} Q^2 = \operatorname{dom} H^{1/2},$$
$$\{Q^1,Q^2\} = 0, \quad \{Q^1,K\} = 0, \quad K = N_0 - N_1, \tag{6.1}$$

where the anticommutators hold in the sense of forms, and N_0 (N_1) project onto even (odd) states.

For the construction of the corresponding supergroup, the Hilbert space $\mathscr{H}_f$ is extended to the $\mathscr{D}_2$ module $\mathscr{H}_f(\Theta_1,\Theta_2)$ with elements

$$\begin{aligned}&\Psi_0 + \Theta_1\Psi_1 + \Theta_2\Psi_2 + \Theta_1\Theta_2\Psi_{12} \in \mathscr{H}_f(\Theta_1,\Theta_2),\\ &\Psi_0,\Psi_1,\Psi_2,\Psi_{12} \in \mathscr{H}_f.\end{aligned} \tag{6.2}$$

The Grassmann algebra of polynomials in the anticommuting parameters Θ_1, Θ_2 with complex coefficients will be denoted by $\mathscr{D}_2$ and the tensor product of $\mathscr{D}_2$ with the Clifford algebra K_{2f} is constructed. The rules for adjointness and the scalar product[6] imply, for example,[9]

$$\begin{aligned}&\langle\Phi|(c_1\Theta_1 + c_2\Theta_2 + c_{12}\Theta_1\Theta_2)\Psi\rangle\\ &\quad = (c_1\Theta_1 + c_2\Theta_2)\langle\Phi|K\Psi\rangle + c_{12}\Theta_1\Theta_2\langle\Phi|\Psi\rangle,\\ &\qquad c_1,c_2,c_{12}\in\mathbb{C}, \quad \Phi,\Psi\in\mathscr{H}_f.\end{aligned} \tag{6.3}$$

With the notation

$$\begin{aligned}&Q = (Q^1 + iQ^2)/\sqrt{2}, \quad \Theta = \Theta_1 + i\Theta_2,\\ &\Theta^* = \Theta_1 - i\Theta_2, \quad \Theta^2 = \Theta^{*2} = 0,\end{aligned} \tag{6.4}$$

one writes the general element of the supergroup generated by $S(2)$ as

$$\begin{aligned}g(t;s,r) &= \exp(itH + isQ\Theta + is^*\Theta^*Q^\dagger + irH\Theta\Theta^*)\\ &= e^{itH}(I + isQ\Theta + is^*\Theta^*Q^\dagger\\ &\quad - \tfrac{1}{2}|s|^2\{Q\Theta,\Theta^*Q^\dagger\} + irH\Theta\Theta^*)\\ &= g(t;0,0)g(0;s,0)g(0;0,r), \quad t,r\in\mathbb{R}, \quad s\in\mathbb{C},\end{aligned} \tag{6.5}$$

which is defined as a sesquilinear form on $\operatorname{dom} Q = \operatorname{dom} H^{1/2}$.

The corresponding composition law becomes

$$g(t;s,r)g(t';s',r') = g(t + t';s + s',r + r' + 2\operatorname{Im}(s's^*)); \tag{6.6}$$

these transformations are superunitary in the sense that

$$g(t;s,r)^\dagger = g(t;s,r)^{-1} = g(-t;-s,-r) \tag{6.7}$$

on form $\operatorname{dom} g(t;s,r) = \operatorname{dom} H^{1/2}$ for t and r real, and s complex.

With the help of the supercharge

$$Q = \sqrt{2}\sum_{k=1}^{f}\varepsilon_k B_k$$

for the f-dimensional fermionic oscillator one obtains the (anti-) commutation relations

$$\begin{aligned}&\{Q,\partial_k\} \subsetneqq \sqrt{2}B_k, \quad \{Q^\dagger,\partial_k\} = 0, \quad [H,\partial_k] \subsetneqq -\partial_k,\\ &[Q^\dagger,B_k] \subsetneqq \sqrt{2}\partial_k, \quad [Q,B_k] = 0, \quad [H,B_k] \subsetneqq -B_k,\\ &\qquad k = 1,\dots,f.\end{aligned} \tag{6.8}$$

Relations (6.8) lead to the following transformation of the fermionic oscillator Lie superalgebra:

$$\begin{aligned}&e^{itH}\partial_k e^{-itH} = e^{-it}\partial_k, \quad e^{itH}B_k e^{-itH} = e^{-it}B_k,\\ &e^{irH\Theta\Theta^*}\partial_k e^{-irH\Theta\Theta^*} \subsetneqq \partial_k - ir\Theta\Theta^*\partial_k,\\ &e^{irH\Theta\Theta^*}B_k e^{-irH\Theta\Theta^*} \subsetneqq B_k - ir\Theta\Theta^*B_k,\\ &g(0;s,0)\partial_k g(0;-s,0) \subsetneqq \partial_k - is\sqrt{2}\Theta B_k - |s|^2\Theta\Theta^*\partial_k,\\ &g(0;s,0)B_k g(0;-s,0) \subsetneqq B_k - is^*\sqrt{2}\Theta^*\partial_k + |s|^2\Theta\Theta^*B_k,\\ &\qquad t,r\in\mathbb{R}, \quad s\in\mathbb{C}, \quad k = 1,\dots,f.\end{aligned} \tag{6.9}$$

An obvious composition of these superunitary transformations leads to an interesting group of automorphisms of the C(A)CR.

The question of whether any automorphism of the C(A)CR with two or more anticommuting parameters is implemented by a unique superunitary transformation leads to rather complicated generalizations of the Lie superalgebra (A3). We consider first the case of one bosonic and one fermionic degree of freedom, and then generalize to $f\geqslant 2$ fermions.

To start the explicit construction of superunitary transformations g for $f = 1$, we define

$$\begin{aligned}g &= \exp(\Theta_1A_1 + \Theta_2A_2 + \Theta_1\Theta_2T)\\ &= I + \Theta_1A_1 + \Theta_2A_2 + \tfrac{1}{2}\Theta_1\Theta_2[A_2,A_1] + \Theta_1\Theta_2T,\\ A_k^\dagger &= A_k, \quad k = 1,2, \quad T^\dagger = T,\\ g^\dagger &\supseteqq \exp(-\Theta_1A_1 - \Theta_2A_2 - \Theta_1\Theta_2T),\end{aligned} \tag{6.10}$$

where A is an odd and T an even operator in the Hilbert space $\mathscr{L}_1 = L^2(d^1x) \otimes \mathscr{G}_1$; then

$$\begin{aligned}\operatorname{dom} g^\dagger &\supseteqq \operatorname{dom} g\\ &= \operatorname{dom} gg^\dagger = \operatorname{dom} g^\dagger g\\ &= \operatorname{dom}[A_2,A_1] \cap \operatorname{dom} T \oplus \Theta_1 \operatorname{dom} A_2\\ &\quad \oplus \Theta_2 \operatorname{dom} A_1 \oplus \Theta_1\Theta_2\mathscr{L}_1;\end{aligned}$$

in addition $gg^\dagger = g^\dagger g = I_{|\operatorname{dom} g}$. The transformed operators $gBg^\dagger$, $gB^\dagger g^\dagger$, $g\partial g^\dagger$, $g\varepsilon g^\dagger$ fulfill again the C(A)CR on appropriately restricted domains.

With the special ansatz

$$\begin{aligned}&A_k = G_k\varepsilon + G_k^\dagger\partial, \quad G_k \text{ closed}, \quad \operatorname{dom} G_k = \operatorname{dom} G_k^\dagger,\\ &G_k = \{A_k,\partial\}, \quad k = 1,2,\end{aligned} \tag{6.11}$$

one obtains the transformed operators

$$\begin{aligned}gBg^\dagger &= B + \Theta_1[A_1,B] + \Theta_2[A_2,B]\\ &\quad + \Theta_1\Theta_2(A_1BA_2 - A_2BA_1\\ &\quad + \tfrac{1}{2}\{[A_2,A_1],B\} + [T,B]),\end{aligned} \tag{6.12}$$

$$\begin{aligned} g\partial g^\dagger &= \partial + \Theta_1\{A_1,\partial\} + \Theta_2\{A_2,\partial\} + \Theta_1\Theta_2(A_2\partial A_1 \\ &\quad - A_1\partial A_2 + \tfrac{1}{2}\{\partial,[A_2,A_1]\} + [T,\partial]) \\ &\supsetneqq \partial + \Theta_1 G_1 + \Theta_2 G_2 + \Theta_1\Theta_2\varepsilon[G_2,G_1] \\ &\quad + \Theta_1\Theta_2\partial(\tfrac{1}{2}[G_2,G_1^\dagger] \\ &\quad - [G_1,G_2^\dagger]) + T'' - T'), \end{aligned}$$

where we decomposed $T = T'\varepsilon\,\partial + \mathrm{T}''\,\partial\varepsilon$ with T' and T'' being self-adjoint.

Conversely, given an automorphism of the C(A)CR, there exists under suitable domain conditions a superunitary transformation implementing it.

The following theorem generalizes this implementation to the case of f fermionic and f bosonic degrees of freedom.

Theorem 4: Let C_k be bounded operators in the separable Hilbert space $\mathscr{H} = \mathscr{H}^0 \oplus \mathscr{H}^1$, and let B_k, D_k^i, G_k^i, E_k, and F_k be densely defined closed operators in $\mathscr{H}$. Assume that B_k, G_k^i, and E_k are even, C_k, D_k^i, and F_k with $i = 1,2$, $k = 1,\ldots,f$, are odd. Define

$$\begin{aligned} &B_k' = B_k + \Theta_1 D_k^1 + \Theta_2 D_k^2 + \Theta_1\Theta_2 E_k, \\ &C_k' = C_k + \Theta_1 G_k^1 + \Theta_2 G_k^2 + \Theta_1\Theta_2 F_k, \\ &B_k'^\dagger = B_k^\dagger - \Theta_1 D_k^{1\dagger} - \Theta_2 D_k^{2\dagger} - \Theta_1\Theta_2 E_k^\dagger, \\ &B_k'^{\dagger\dagger} = B_k', \\ &C_k'^\dagger = C_k^\dagger + \Theta_1 G_k^{1\dagger} + \Theta_2 G_k^{2\dagger} - \Theta_1\Theta_2 F_k^\dagger, \\ &C_k'^{\dagger\dagger} = C_k', \quad k = 1,\ldots,f. \end{aligned} \tag{6.13}$$

Assume that these generalized operators fulfill the C(A)CR in the sense of operator polynomials,

$$\begin{aligned} &\{C_k',C_j'\} = 0, \quad \{C_k',C_j'^\dagger\} \subseteq \delta_{kj} I, \\ &[B_k',B_j'] = 0, \quad [B_k',B_j'^\dagger] \subseteq \delta_{kj} I, \\ &[B_k',C_j'] = 0, \quad [B_k',C_j'^\dagger] = 0, \quad k,j = 1,\ldots,f. \end{aligned} \tag{6.14}$$

Moreover, B_k and $C_k^\dagger$ are supposed to obey the same conditions as in Theorem 3 and can therefore be identified with $(x_k + \partial/\partial x_k)/\sqrt{2}$ and ε_k, $k = 1,\ldots,f$, and $\mathscr{H}$ can be identified with $\mathscr{L}_f$.

Next we introduce components for

$$G_k^i = \sum_{\substack{1 \leqslant p_1 < \cdots < p_r \leqslant f \\ 1 \leqslant q_1 < \cdots < q_s \leqslant f \\ r+s = 0,2,4,\ldots}} G^i_{k,p_1\cdots p_r q_1\cdots q_s} \varepsilon_{p_1}\cdots\varepsilon_{p_r}\partial_{q_1}\cdots\partial_{q_s},$$

$$i = 1,2, \quad k = 1,\ldots,f, \tag{6.15}$$

and similar components $D^i_{k,p_1\cdots p_r q_1\cdots q_s}$, with $r+s$ odd, for D_k^i components for E_k with $r+s$ even, and components for F_k with $r+s$ odd; assume that $\mathscr{C}$ is an invariant dense domain for all these components and their adjoints as well as for B_k and $B_k^\dagger$ with $k = 1,\ldots,f$.

Then there exist unique symmetric odd operators $\dot{A}_1,\dot{A}_2$ and a symmetric even operator $\dot{T}$, defined on the domain $\mathscr{C}_f = \mathscr{C} \otimes \mathscr{G}_f$, such that

$$\begin{aligned} &G_k^i \supseteq \{\dot{A}_i,\partial_k\}, \quad D_k^i \supseteq [\dot{A}_i,B_k], \quad i = 1,2, \\ &F_k \supseteq \dot{A}_2\,\partial_k\dot{A}_1 - \dot{A}_1\partial_k\dot{A}_2 + \tfrac{1}{2}\{\partial_k,[\dot{A}_2,\dot{A}_1]\} + [\dot{T},\partial_k], \\ &E_k \supseteq \dot{A}_1 B_k\dot{A}_2 - \dot{A}_2 B_k\dot{A}_1 + \tfrac{1}{2}\{B_k,[\dot{A}_2,\dot{A}_1]\} + [\dot{T},B_k], \\ &\qquad k = 1,\ldots,f. \end{aligned} \tag{6.16}$$

The proof of this theorem is given in Appendix B.

Corollary 5: If $\ddot{T}$ is another symmetric even operator on $\mathscr{C}_f$ that fulfills (6.16), too, instead of $\dot{T}$, and if each component of $\dot{T} - \ddot{T}$ is essentially self-adjoint on $\mathscr{C} = S(\mathbb{R}^f)$, then $\dot{T} - \ddot{T} \subseteq tI$ for some real t.

The proof of Corollary 5 follows from Kato's condition,[14] since $(x_k + iI)\mathscr{C} = (p_k + iI)\mathscr{C} = \mathscr{C}$, which implies that $\dot{T} - \ddot{T}$ commutes strongly with both x_k and p_k for $k = 1,\ldots,f$.

Corollary 6: Under the conditions of Theorem 4 it follows that the transformation

$$\begin{aligned} \dot{g} &= \exp(\Theta_1\dot{A}_1 + \Theta_2\dot{A}_2 + \Theta_1\Theta_2\dot{T}) \\ &= I + \Theta_1\dot{A}_1 + \Theta_2\dot{A}_2 + \tfrac{1}{2}\Theta_1\Theta_2[\dot{A}_2,\dot{A}_1] + \Theta_1\Theta_2\dot{T}, \\ \dot{g}^\dagger &= I - \Theta_1\dot{A}_1^\dagger - \Theta_2\dot{A}_2^\dagger - \tfrac{1}{2}\Theta_1\Theta_2[\dot{A}_2,\dot{A}_1]^\dagger - \Theta_1\Theta_2\dot{T}^\dagger \\ &\supseteq \exp(-\Theta_1\dot{A}_1^\dagger - \Theta_2\dot{A}_2^\dagger - \Theta_1\Theta_2\dot{T}^\dagger), \end{aligned} \tag{6.17}$$

with $\dot{g}\dot{g}^\dagger = \dot{g}^\dagger\dot{g} = I_{|\mathrm{dom}\,\dot{g}}$, $\mathrm{dom}\,\dot{g}^\dagger \supseteq \mathrm{dom}\,\dot{g} = \mathrm{dom}\,\dot{g}^\dagger\dot{g} = \mathrm{dom}\,\dot{g}\dot{g}^\dagger$, implements the automorphism (6.13), i.e.,

$$\mathrm{B}_k' \supseteq \dot{g}B_k\dot{g}^\dagger, \quad C_k' \supseteq \dot{g}\partial_k\dot{g}^\dagger, \quad k = 1,\ldots,f. \tag{6.18}$$

Corollary 7: For the case $f = 1$, let

$$\begin{aligned} &G_1^i = G_1^{i\prime} I + G_1^{i\prime\prime}\varepsilon_1\,\partial_1, \\ &\mathrm{dom}\,G_1^{i\prime} = \mathrm{dom}\,G_1^{i\prime\dagger} \subseteq \mathrm{dom}\,G_1^{i\prime\prime}, \quad i = 1,2; \end{aligned} \tag{6.19}$$

assume that $\mathscr{C}$ is a core for $G_1^{i\prime}$. Then one especially obtains that

$$G_1^i = \{A_i,\partial_1\}, \quad G_1^{i\prime\prime} = 0, \quad A_i = G_1^{i\prime}\varepsilon_1 + G_1^{i\prime\dagger}\partial_1 = A_i^\dagger,$$
$$i = 1,2. \tag{6.20}$$

Example 4: The choice

$$\begin{aligned} A_1 &= i\sqrt{2}\sum_{k=1}^{f}(-sB_k\varepsilon_k + s^*B_k^\dagger\partial_k), \\ A_2 &= \sqrt{2}\sum_{k=1}^{f}(sB_k\varepsilon_k + s^*B_k^\dagger\partial_k), \quad s\in\mathbb{C}, \end{aligned} \tag{6.21}$$

corresponds to self-adjoint supercharges of the f-dimensional fermionic oscillator. Using its Hamilton operator

$$T = 2r\sum_{k=1}^{f}(B_k^\dagger B_k + \varepsilon_k\partial_k), \quad r\in\mathbb{R}, \tag{6.22}$$

yields the supergroup element $g(0;s,r)$, which was defined in (6.5), for this model, and which implements the automorphism group (6.9).

ACKNOWLEDGMENT

L. P. was supported in part by "Fonds zur Förderung der wissenschaftlichen Forschung in Österreich," project No. P5588.

APPENDIX A: PROOF OF THEOREM 3

Here we give the proof of Theorem 3. An appropriate operator basis for $\mathscr{G}_f$ is given by the 2^{2f} monomials

$$\begin{aligned} &I, \quad \varepsilon_{p_1}\cdots\varepsilon_{p_r} \quad (1 \leqslant p_1 < \cdots < p_r \leqslant f), \\ &\partial_{q_1}\cdots\partial_{q_s} \quad (1 \leqslant q_1 < \cdots < q_s \leqslant f), \quad \varepsilon_{p_1}\cdots\varepsilon_{p_r}\partial_{q_1}\cdots\partial_{q_s}. \end{aligned} \tag{A1}$$

The (anti-) commutation relations

$$\{\varepsilon_{p_1}\cdots\varepsilon_{p_r}\partial_{q_1}\cdots\partial_{q_s},\partial_k\}=\begin{cases}0, & \text{for } k\notin\{p_1\cdots p_r\},\\ (-)^{i-1}\varepsilon_{p_1}\cdots\not\varepsilon_{p_i}\cdots\varepsilon_{p_r}\partial_{q_1}\cdots\partial_{q_s}, & \text{for } k=p_i,\ r+s \text{ odd},\end{cases} \tag{A2}$$

$$[\varepsilon_{p_1}\cdots\varepsilon_{p_r}\partial_{q_1}\cdots\partial_{q_s},\partial_k]=(-1)^i\varepsilon_{p_1}\cdots\not\varepsilon_{p_i}\cdots\varepsilon_{p_r}\partial_{q_1}\cdots\partial_{q_s},\quad \text{for } k=p_i,\quad r+s \text{ even},$$

follow easily from (2.4). The C(A)CR (5.16) imply the Lie superalgebra

$$[D_k,B_l]-[D_l,B_k]=0, \tag{A3a}$$

$$[B_k,G_l]+\{\partial_l,D_k\}=0, \tag{A3b}$$

$$[B_k,G_l^\dagger]+\{\varepsilon_l,D_k\}=0, \tag{A3c}$$

$$[G_k,\partial_l]+[G_l,\partial_k]=0, \tag{A3d}$$

$$[G_k,\varepsilon_l]+[G_l^\dagger,\partial_k]=0, \tag{A3e}$$

$$[D_k,B_l^\dagger]+[D_l^\dagger,B_k]=0,\quad k,l=1,\dots,f, \tag{A3f}$$

in the sense of operator polynomials. Next we insert the ansatz (5.19). (A3d) implies for $k=l$ that $G_k^{\{p,q\}}=0$ if $k\in\{p_1\dots p_r\}$; (A3d) implies for $k\neq l$ that

$$\begin{aligned}(-)^l\dot G_k^{p_1\cdots\not p_l\cdots p_r,q_1\cdots q_s}&\\ =(-)^n\dot G_h^{p_1\cdots\not p_n\cdots p_r,q_1\cdots q_s},&\quad k=p_l,\quad h=p_n,\\ r+s-1=0,2,4,\dots,&\end{aligned} \tag{A4}$$

where $\dot G_k^{\{p,q\}}$ denotes the restriction of the components defined by (5.19) to $\mathscr{C}$. We therefore may define

$$\begin{aligned}A'=&\sum_{\substack{1\leqslant p_1<\cdots<p_r\leqslant f\\ r\ \text{odd}}}\varepsilon_{p_1}\cdots\varepsilon_{p_r}A'^{p_1\cdots p_r}\\ &+\sum_{\substack{1\leqslant p_1<\cdots<p_r\leqslant f\\ 1\leqslant q_1<\cdots<q_s\leqslant f\\ r+s\ \text{odd}}}\varepsilon_{p_1}\cdots\varepsilon_{p_r}\partial_{q_1}\cdots\partial_{q_s}A'^{p_1\cdots p_r q_1\cdots q_s},\end{aligned} \tag{A5}$$

on dom $A'=\mathscr{C}_f$, with

$$\begin{aligned}&A'^{p_1\cdots p_r}=(-)^{l-1}\dot G_k^{p_1\cdots\not p_l\cdots p_r},\quad p_l=k,\quad 1\leqslant l\leqslant r,\\ &A'^k=\dot G_k^0,\quad k=1,\dots,f,\\ &A'^{p_1\cdots p_r,q_1\cdots q_s}=(-)^{l-1}\dot G_k^{p_1\cdots\not p_l\cdots p_r,q_1\cdots q_s},\end{aligned} \tag{A6}$$

and obtain

$$\begin{aligned}\dot G_k=&\sum_{\substack{\{p,q\}\\ r+s=0,2,4,\dots}}\varepsilon_{p_1}\cdots\varepsilon_{p_r}\partial_{q_1}\cdots\partial_{q_s}\dot G_k^{\{p,q\}}\\ &=\{A',\partial_k\},\quad k=1,\dots,f.\end{aligned} \tag{A7}$$

Define next

$$\begin{aligned}&\dot H_k=[B_k,A']+\dot D_k,\quad \dot K_k=[B_k,A'^\dagger{}_{|\mathscr{C}_f}]+\dot D_k,\\ &\dot D_k=D_{k|\mathscr{C}_f},\quad \text{dom}\,\dot H_k=\text{dom}\,\dot K_k=\mathscr{C}_f,\quad k=1,\dots,f,\end{aligned} \tag{A8}$$

using (5.21); (A3b) and (A3c) imply then that $\dot H_k$ does not contain components with $\varepsilon_{p_1}\cdots\varepsilon_{p_r}$, and $\dot K_k$ does not contain components with $\partial_{q_1}\cdots\partial_{q_s}$. Define $F'=A'-A'^\dagger$ on dom $F'=\mathscr{C}_f$; using (A8) one obtains

$$[B_k,F'^{\{p,q\}}]=[B_k^\dagger,F'^{\{p,q\}}]=0, \tag{A9}$$

for $\{p,q\}=\{p_1\cdots p_r,q_1\cdots q_s\}$ with $rs\geqslant 1$, $r+s$ odd; here the $F'^{\{p,q\}}$ denote the components of F' in the basis (A1). Redefining

$$\dot A=A'+\sum_{\substack{1\leqslant q_1<\cdots<q_s\leqslant f\\ s\ \text{odd}}}\partial_{q_s}\cdots\partial_{q_1}A'^{q_1\cdots q_s\dagger},\quad \text{dom}\,\dot A=\mathscr{C}_f, \tag{A10}$$

one obtains the desired (anti-) commutators

$$[B_k,\dot A]=-\dot D_k,\quad \{\partial_k,\dot A\}=\dot G_k,\quad k=1,\dots,f. \tag{A11}$$

The redefinition (A10) implies next that

$$\begin{aligned}\dot F=\dot A-\dot A^\dagger=&\sum_{\substack{1\leqslant p_1<\cdots<p_r\leqslant f\\ 1\leqslant q_1<\cdots<q_s\leqslant f\\ rs\geqslant 1,r+s\ \text{odd}}}\varepsilon_{p_1}\cdots\varepsilon_{p_r}\partial_{q_1}\cdots\partial_{q_s}\dot F^{\{p,q\}},\\ &\text{dom}\,\dot F=\mathscr{C}_f,\end{aligned} \tag{A12}$$

and obviously $\dot F\subseteq-\dot F^\dagger$. As the last step of the proof we insert (A11) into (A3e) and calculate

$$\begin{aligned}\dot F\partial_k\varepsilon_l+\partial_k\varepsilon_l\dot F+\partial_k\dot F\varepsilon_l-\varepsilon_l\dot F\partial_k-\delta_{kl}\dot F=0,&\\ k,l=1,\dots,f.&\end{aligned} \tag{A13}$$

Inserting (A12) into (A13) gives $\dot F=0$, which finally implies that $\dot A$ is symmetric. This symmetry of $\dot A$ obviously implies its uniqueness.

APPENDIX B: PROOF OF THEOREM 4

Here we give the proof of Theorem 4. Inserting (6.13) into the C(A)CR (6.14) one obtains super-commutation relations for the closed operators involved. These equations, which do not contain E_k and F_k, $k=1,\dots,f$, are just the same as in (A3) and lead therefore to the symmetric odd operators $\dot A_i$. The remaining relations look at a first glance rather complicated, but can be simplified after introducing

$$\begin{aligned}&\widetilde F_k=F_k-\dot A_2\partial_k\dot A_1+\dot A_1\partial_k\dot A_2-\tfrac12\{\partial_k,[\dot A_2,\dot A_1]\},\\ &\widetilde E_k=E_k-\dot A_1B_k\dot A_2+\dot A_2B_k\dot A_1-\tfrac12\{B_k,[\dot A_2,\dot A_1]\},\\ &\qquad k=1,\dots,f,\end{aligned} \tag{B1}$$

and lead to the Lie superalgebra

$$\{\partial_k,\widetilde F_l\}+\{\partial_l,\widetilde F_k\}=0, \tag{B2a}$$

$$[B_k,\widetilde F_l]+[\widetilde E_k,\partial_l]=0, \tag{B2b}$$

$$[B_k,\widetilde F_l^\dagger]+[\varepsilon_l,\widetilde E_k]=0, \tag{B2c}$$

$$[B_k,\widetilde E_l]-[B_l,\widetilde E_k]=0, \tag{B2d}$$

$$[B_k,\widetilde E_l^\dagger]+[B_l^\dagger,\widetilde E_k]=0, \tag{B2e}$$

$$\{\varepsilon_k,\widetilde F_l\}-\{\partial_l,\widetilde F_l^\dagger\}=0,\quad k,l=1,\dots,f. \tag{B2f}$$

Now, following the analysis after Eqs. (A3) of Appendix A one obtains an operator $\dot T$ of the form

$$\dot{T} = \sum_{\substack{k=p_l \\ r+s \text{ even} \\ r>1}} (-)^l \widetilde{F}_{k,p_1\cdots\not{p}_l\cdots p_r\{q\}} \varepsilon_{p_1}\cdots\varepsilon_{p_r}\partial_{q_1}\cdots\partial_{q_s}$$

$$+ \sum_{\substack{k=p_l \\ r>1}} (-1)^l \widetilde{F}^{\dagger}_{k,p_1\cdots\not{p}_l\cdots p_r}\partial_{p_r}\cdots\partial_{p_1} \tag{B3}$$

in analogy to (A10); $\dot{T}$ fulfills the commutation relations

$$\widetilde{F}_k = [\dot{T},\partial_k], \quad \widetilde{E}_k = [\dot{T},B_k], \quad k = 1,\ldots,f. \tag{B4}$$

If one inserts (B3) into (B2f) one finds that T is symmetric.

[1] J. Wess and B. Zumino, Nucl. Phys. B **70**, 39 (1974).

[2] E. Witten, Nucl. Phys. B **185**, 513, (1981).

[3] P. Salomonson and J. W. van Holten, Nucl. Phys. B **196**, 509 (1982).

[4] M. de Crombrugghe and V. Rittenberg, Ann. Phys. (NY) **151**, 99 (1983).

[5] A. Arai, J. Funct. Anal. **60**, 378 (1985).

[6] H. Grosse and L. Pittner, J. Phys. A **20**, 4265 (1987).

[7] C. R. Putnam, *Commutation Properties of Hilbert Space Operators and Related Topics* (Springer, Berlin, 1967).

[8] F. A. Berezin, *The Method of Second Quantization* (Academic, New York, 1966).

[9] L. van Hove, Nucl. Phys. B **207**, 15 (1982).

[10] T. Kato, *Perturbation Theory for Linear Operators* (Springer, New York, 1980).

[11] H. Nicolai, J. Phys. A **9**, 1497 (1976).

[12] M. Scheunert, "The theory of Lie superalgebras," *Lecture Notes in Mathematics*, Vol. 716 (Springer, Berlin, 1979).

[13] N. Jacobson, *Basic Algebra I* (Freeman, San Francisco, 1974).

[14] P. E. Jørgensen and R. T. Moore, *Operator Commutation Relations* (Reidel, Dordrecht, 1984), p. 287.

Approximate eigenvalues of parameter-dependent systems from boundaries of level sets

Gustavo A. Arteca
Department of Chemistry, University of Saskatchewan, Saskatoon, Saskatchewan, Canada S7N 0W0

Paul G. Mezey
Departments of Chemistry and Mathematics, University of Saskatchewan, Saskatoon, Saskatchewan, Canada S7N 0W0

(Received 3 June 1987; accepted for publication 31 August 1987)

Some basic properties of the energy level sets for parameter-dependent systems are analyzed. A class of Hamiltonians depending linearly on two parameters is considered as a working example. It is shown that an approximate computation of the boundary of the level set can be useful to derive analytical representations of the eigenvalues of the system. The expressions deduced for the eigenenergies have the correct analytic behavior in the whole range of all physical parameters. In the simplest case, a nonvariational upper bound to the ground state energy is derived that satisfies both virial and Hellmann–Feynman theorems, as well as first-order perturbation theory. Furthermore, some approximations are obtained, within the framework of the scaling variational method, that are numerical upper and lower bounds to the exact eigenvalues. Applications are illustrated with a family of anharmonic oscillators.

I. INTRODUCTION

The derivation of approximate expressions for eigenvalues of parameter-dependent systems has attracted attention time and again. Several developments have been motivated by the desire to determine which is the basic information that should be used to describe, qualitatively, the analytic behavior of the energy in such systems.[1–9]

It is well known that the virial theorem (VT) and Hellmann–Feynman theorem (HFT) lead to a differential equation that fixes the essential dependence of the eigenvalues on the parameters contained in the Hamiltonian. In particular, when an appropriate general trial function is optimized variationally, the approximate expression for the energy satisfies both theorems and it allows one to describe not only the correct qualitative behavior of *all* the eigenvalues with respect to the above parameters,[1–9] but also in terms of the quantum numbers.[3,10] This constitutes the essence of the so-called scaling variational method (SVM).[3,11–14] The analytic description it provides for the eigenvalues is believed to be independent of the basis functions chosen. However, this has not yet been proved.[10]

Recently, several authors have proposed a more general approximation to the problem above. It consists basically in building families of functionals that are solutions of the differential equation determined by the HFT and VT.[1–5,9,15,16] These functionals include as a particular case the approximation given by the SVM, but they can be easily generalized.[15] A very attractive feature of this approximation is that it allows one to take into account all the information available about the systems, while it always keeps the correct dependence of the energy on all the parameters of interest. Among the usual sources of such complementary information one should mention the Rayleigh–Schrödinger perturbation theory (RSPT) and the semiclassical limit of the energy.[5,16] This approximation in terms of functionals can explain easily some intriguing results, for instance, the cause of the coincidence between the analytic behavior predicted by the JWKB method and the variational theorem.[17]

As a result of the above studies, there is some interest in searching for new, simple representations for the energy based on different principles. Of course, the new expressions should present, if possible, the convenient features of the functional solutions of the HFT–VT differential equation, as well as leading to possible improvements.

One of the aims of this paper is to construct such a family of approximate expressions for the eigenvalues of parameter-dependent systems.

Our starting point are some relationships derived previously from level sets of the energy of systems depending on linear parameters,[18] including, in particular, those for the electronic energy of polyatomic molecules in the abstract space of nuclear charges.[18–23] This space can be given a topological structure from its partitioning into level sets, and this approach has been useful for deducing different bounds for the electronic energy.[18–21] Recently, we have shown that, in the simplest case of diatomic molecules, the curvature properties of the boundaries of the above level sets can be used to obtain approximate analytical expressions (upper bounds) for the electronic energy as a function of both nuclear charges and interatomic distances.[22,23] It was also shown that the basic structure of the function providing the bounds was the same as it is for the exact electronic energy.[23] However, it was not clear if those bounds had any relationship with the ones provided by the usual variational method, and if they could be improved and generalized to some other systems.

In this paper a method to derive approximate expressions for eigenvalues is presented. In its formulation we make use of the basic properties of the boundaries of level sets. For the sake of simplicity, we confine ourselves to systems depending linearly on two parameters; in this case all the equations posed by the method can be handled rather

0022-2488/88/010119-09$02.50

easily. In spite of their simplicity, these systems include some models of actual physical interest. Moreover, the extension to other problems depending on more parameters, as well as those presenting nonlinear dependences, is possible.

In the case of two-parameter systems, the boundaries of level sets are simple curves, which could be termed "constant energy trajectories" (CET).[21] From now on, we refer to the procedure as the CET method. As it is shown below, the technique allows one to derive new approximate expressions for the eigenvalues that share many of the properties characteristic of the functional solutions of the HFT–VT differential equation. As in previous methods, the algorithm seems well adapted to introduce available analytical information. As is shown in this paper, in many cases the present alternative has some advantages.

The paper is organized as follows. In Sec. II we summarize briefly the basic properties of the CET's for a general, well-defined family of Hamiltonians. A very simple nonvariational upper bound to the eigenvalues is derived in this section. The analytical properties of this bound, as well as some possible methods for its generalization, are also discussed in Sec. II. Following the ideas presented in Sec. II, in Sec. III we derive several approximate expressions for the eigenvalues of anharmonic oscillators, in order to illustrate how the method works. Numerical results are shown for some eigenstates of the quartic anharmonic oscillator. A discussion comparing the procedure with some other techniques involving similar information is also included. We conclude in Sec. IV with some comments about possible extensions of the method.

II. BASIC RESULTS AND ELEMENTARY BOUNDS FOR EIGENVALUES

In this paper we deal with the following family of differential operators:

$$H(Z_1,Z_2) = -\Delta + Z_1V_1(\mathbf{x}) + Z_2V_2(\mathbf{x}), \quad \mathbf{x}\in{}^s\mathbb{R}, \quad s\geqslant 1, \tag{1}$$

where Δ is the Laplacian operator in ${}^s\mathbb{R}$. The potential in (1) is defined according to the following conditions.

(i) Z_1 and Z_2 are positive real parameters: $Z_1\in\mathbb{R}_0^+$, $Z_2\in\mathbb{R}_0^+$, where we use the notation $\mathbb{R}_0^+$ for the set of non-negative real numbers.

(ii) If there exists some $\mathbf{x}_0\in{}^s\mathbb{R}(\|\mathbf{x}_0\| < \infty)$ so that $V_i(\mathbf{x}_0) = 0$, then $\|\nabla V_i(\mathbf{x})\|(\mathbf{x} = \mathbf{x}_0) = 0$, and the Hessian matrix $(\nabla\nabla V_i(\mathbf{x}))(\mathbf{x} = \mathbf{x}_0)$ has either only negative or only positive eigenvalues, for both $i = 1$ and $i = 2$.

(iii) $\operatorname{sgn}(V_1(\mathbf{x})) = \operatorname{sgn}(V_2(\mathbf{x}))$, for all $\mathbf{x}\in{}^s\mathbb{R}$.

We include a further assumption that will allow us to advance farther in the analytical derivation (even though it is not essential to develop the method).

(iv) $V_1(\mathbf{x})$ and $V_2(\mathbf{x})$ are homogeneous functions of degree N_1 and N_2, respectively.

Note that in Sec. IV the changes in the formulation, required by a relaxation in some of above conditions, are discussed briefly.

In order to define the level sets (and their boundaries) corresponding to a given eigenstate, the spectrum of $H(Z_1,Z_2)$ must fulfill certain properties under continuous variation of Z_1 and Z_2. Namely, we assume the following two properties.

(v) There exists a nonempty set D, $D\subset\mathbb{R}_0^+\times\mathbb{R}_0^+$, so that for all $(Z_1,Z_2)\in D$, $H(Z_1,Z_2)$ has a nonempty set of discrete eigenstates.

(vi) $D\cap\mathbb{R}_0^+\times\{0\}\neq\varnothing$ and $D\cap\{0\}\times\mathbb{R}_0^+\neq\varnothing$, where $\varnothing$ is the empty set. That is, the Hamiltonians $H(Z_1,0)$ and $H(0,Z_2)$ have spectra that can be obtained from that of $H(Z_1,Z_2)$ in the limit $Z_2\to 0$ and $Z_1\to 0$, respectively. [Notice that $(0,0)\notin D$.] In order to simplify the analysis that follows, we will assume the discrete spectrum to be nondegenerate.

The above properties (iii), (v), and (vi) guarantee the existence of a discrete spectrum

$$\{E_m(Z_1,Z_2),\ m = 0,1,\ldots,N;\ N\geqslant 0\}$$

over variations of both Z_1 and Z_2 within a certain range. Furthermore, according to the HFT, these eigenvalues are monotonic functions of both Z_1 and Z_2:

$$\begin{aligned}\left(\frac{\partial E_m(Z_1,Z_2)}{\partial Z_1}\right)_{Z_2} &= \{\langle \mathrm{V}_1\rangle_m\}_{(Z_1,Z_2)},\\ \left(\frac{\partial\,\mathrm{E}_m(Z_1,Z_2)}{\partial Z_2}\right)_{Z_1} &= \{\langle V_2\rangle_m\}_{(Z_1,Z_2)}.\end{aligned} \tag{2}$$

Here m must be understood as the set of good quantum numbers necessary to specify completely the spectrum. In the above equations, and in what follows, the symbol $\{\cdots\}_{(Z_1,Z_2)}$ stands for the expectation values calculated by integrating with an eigenstate $\psi_m(Z_1,Z_2;\mathbf{x})$ of Hamiltonian (1) in the integrand, whereas the notation $\langle V_i\rangle_m$ stands for

$$\langle\psi_m(Z_1Z_2;\mathbf{x})\,|\,V_i(\mathbf{x})\,|\,\psi_m(Z_1,Z_2;\mathbf{x})\rangle,$$

with $i = 1$ or 2.

We are now in the position to introduce the concept of constant energy trajectory (CET). A CET is simply a continuous curve defined on D, so that every point on it represents a pair of values (Z_1',Z_2') for which the mth eigenvalue of $H(Z_1',Z_2')$ has a fixed numerical value. This curve "connects" the spectra of $H(Z_1,0)$ and $H(0,Z_2)$.[21–23]

In order to define the CET we proceed as follows: let $(Z_0,0)$ and $(0,Z_0')$ be two points in D, so that

$$E_m(Z_0,0) = E_m(0,Z_0'). \tag{3}$$

It is easily shown that, according to the assumptions above for the spectrum and the potential, such a number Z_0' exists for a given Z_0. Using the homogeneity properties of the potential and the Symanzik scaling,[24] we deduce the unitary equivalences of Hamiltonians,

$$H(Z_1,0)\cong(Z_1)^{2/(N_1+2)}H(1,0), \tag{4a}$$

$$H(0,Z_2)\cong(Z_2)^{2(N_2+2)}H(0,1), \tag{4b}$$

which lead us to the result

$$Z_0' = \{E_m(1,0)/E_m(0,1)\}^{(N_2+2)/2}(Z_0)^{(N_2+2)/(N_1+2)}. \tag{5}$$

Observe that, as a result of the monotonicity of the energy as a function of Z_1 and Z_2, there exists a single value

$$Z^* = \{E(1,0)/E(0,1)\}^{(N_2+2)(N_1+2)/2(N_2-N_1)}$$

so that $Z_0 = Z_0' = Z^*$, as long as $N_2\neq N_1$. If $N_2 = N_1$, we get $Z_0 = Z_0'$ for any Z_0. In the general case of $N_2 \neq N_1$, the condition to have $Z_0 > Z_0'$ will depend on whether $Z_0 > Z^*$ or $Z_0 < Z^*$, and on the ratio between the mth eigenvalues of the Hamiltonians $H(1,0)$ and $H(0,1)$. This latter property depends, in turn, on the values of N_1 and N_2.

We now introduce the CET by means of a function f_m: $[0,Z_0] \to [0,Z_0']$ defined according to the following properties:

$$\lim_{Z_1 \downarrow 0} f_m(Z_1) = Z_0', \quad \lim_{Z_1 \uparrow Z_0} f_m(Z_1) = 0, \tag{6a}$$

$$E_m(Z_1, f_m(Z_1)) = E_m(Z_0,0), \quad \text{for all } Z_1 \in (0,Z_0). \tag{6b}$$

It is easily proved from the monotonicity properties of the energy, and the existence of Z_0' for every Z_0, that $Z_2 = f_m(Z_1)$ is a bijective function.

We can introduce now a *level set* $F_m(Z_0)$ on D as[18,19]

$$F_m(Z_0) = \{(Z_1,Z_2) \in D\colon\ E_m(Z_1,Z_2) \geqslant E_m(Z_0,0)\}, \tag{7}$$

for which the CET $f_m(Z_1)$ above is the *boundary*[19]:

$$G_m(Z_0) = \{(Z_1,Z_2) \in D\colon\ Z_2 = f_m(Z_1)\}. \tag{8}$$

If $E_m(Z_1,Z_2) > E_m(Z_0,0)$, the condition to have either $Z_2 > f_m(Z_1)$ or $Z_2 < f_m(Z_1)$ depends on the potential. For instance, if $V_i(\mathbf{x}) < 0$, and $N_i > -2$, $i = 1,2$, it follows that for all $(Z_1,Z_2) \in F_m(Z_0)$, $Z_2 \leqslant f_m(Z_1)$. This is typically the case of one-electron diatomic molecules [Coulombic potentials in Eq. (1)].[19,21]

The CET $f_m(Z_1)$ is in general an unknown function. However, we can determine some of its basic properties, which will turn out to be useful for deriving approximate expressions for the eigenvalues. It is worth commenting that the main properties of the level sets (and their boundaries) can be deduced from the use of the variational theorem to approximate the eigenvalues.[18] However, we shall show here that the analysis of the exact CET's can lead to nonvariational approximations to the eigenenergies, with the same analytic structure that is expected for the exact ones.

Formally, the function $f_m(Z_1)$ can be built from expectation values of the potential. To that purpose let us restrict our discussion to the eigenvalue problem of the operator $H(Z_1, f_m(Z_1))$ [where $f_m(Z_1)$ is not yet determined]. According to Eqs. (6), its expectation value with $\psi_m(Z_1, f_m(Z_1);\mathbf{x})$ is a constant for all $Z_1 \in [0,Z_0]$:

$$\frac{\partial \{\langle H(Z_1, f_m(Z_1))\rangle_m\}_{(Z_1,f_m(Z_1))}}{\partial Z_1} = 0. \tag{9}$$

The HFT and Eq. (9) give us the expression for the derivative of function $f_m(Z_1)$, with respect to Z_1:

$$f_m'(Z_1) = -\{\langle V_1\rangle_m / \langle V_2\rangle_m\}_{(Z_1,f_m(Z_1))} < 0, \quad \text{for all } Z_1 \in [0,Z_0]. \tag{10}$$

Using Eq. (6) we can integrate Eq. (10). Consequently, for a given pair of numbers $Z_0, Z_1 \in (0,Z_0)$, there exists a single real value Z_2 so that (Z_1,Z_2) is in $G_m(Z_0)$; this is given by

$$Z_2 = f_m(Z_1) = \int_{Z_1}^{Z_0} \left\{ \frac{\langle V_1\rangle_m}{\langle V_2\rangle_m} \right\}_{(s,f_m(s))} ds \tag{11}$$

$$= Z_0' - \int_0^{Z_1} \left\{ \frac{\langle V_1\rangle_m}{\langle V_2\rangle_m} \right\}_{(s,f_m(s))} ds. \tag{12}$$

Equations (11) and (12) provide an integral equation for the CET. Naturally, for nontrivial problems these equations cannot be solved exactly to obtain $f_m(Z_1)$. However, we shall see below that the integral representation of $f_m(Z_1)$ can be used to obtain approximate expressions for it. In turn, we will discuss how approximations (bounds) to the exact $f_m(Z_1)$ can lead to approximations (bounds) to the eigenvalues.

It is relevant to our discussion below to determine the sign of the second derivative of the function f_m. It can be determined easily using again the HFT. We start from the equality

$$\frac{\partial^2 E_m(Z_1, f_m(Z_1))}{\partial Z_1^2} = 0 = 2\left\langle \frac{\partial \psi_m}{\partial Z_1} \middle| \frac{\partial H(Z_1, f_m(Z_1))}{\partial Z_1} \middle| \psi_m \right\rangle + \left\{ \left\langle \frac{\partial^2 H(Z_1, f_m(Z_1))}{\partial Z_1^2} \right\rangle_m \right\}_{(Z_1,f_m(Z_1))}. \tag{13}$$

Upon expanding the Hamiltonian in Taylor series,

$$H(Z_1 + \delta Z_1, f_m(Z_1 + \delta Z_1)) \approx H(Z_1, f_m(Z_1)) + \delta Z_1 \frac{\partial H(Z_1, f_m(Z_1))}{\partial Z_1} + O((\delta Z_1)^2), \tag{14}$$

the standard nondegenerate RSPT up to first order allows us to compute the derivative of the wave function in Eq. (13):

$$\frac{\partial \psi_m(Z_1, f_m(Z_1))}{\partial Z_1} = \sum_{s=0\ (s\neq m)} [E_m - E_s]^{-1} \times \left\langle \psi_m \middle| \frac{\partial H(Z_1, f_m(Z_1))}{\partial Z_1} \middle| \psi_s \right\rangle \psi_s. \tag{15}$$

Summation in Eq. (15) must be understood as running over all states with the same symmetry as ψ_m, including both discrete and continuum states.

Introducing (15) in (13), and noticing the relationship between the second partial derivative of the Hamiltonian and the second derivative of the function $f_m(Z_1)$, we get

$$f_m''(Z_1) = [2/\{\langle V_2\rangle_m\}_{(Z_1,f_m(Z_1))}] \sum_{s=0\ (s\neq m)} [E_s - E_m]^{-1} \times \left| \left\langle \psi_m \middle| \frac{\partial H(Z_1 f_m(Z_1))}{\partial Z_1} \middle| \psi_s \right\rangle \right|^2. \tag{16}$$

This equation reveals that for the *lowest state* in every manifold of eigenstates of H with distinct symmetry (say $m = M$), the following equality holds:

$$\operatorname{sgn}(f_M''(Z_1)) = \operatorname{sgn}(V_2(\mathbf{x})), \tag{17}$$

according to the assumptions already commented about the sign of the potential. Observe that Eq. (17) establishes, for example, the concavity from below of the boundaries of level sets of Hamiltonians with Coulombic potentials.[19–21]

The above results for the first and second derivatives of the function $f_m(Z_1)$ lead us to an important conclusion in terms of the ratio of expectation values of V_1 and V_2. This result is condensed into the following lemma.

Lemma 1: Let M be a quantum number or a set of quantum numbers standing for the lowest eigenstate of a given manifold of eigenstates of the Hamiltonian (1), defined according to properties (i)–(vi). Then (a) if $V_2(\mathbf{x})<0$, then

$$\max\{\langle V_1\rangle_M/\langle V_2\rangle_M\}_{(Z_1,f_M(Z_1))} = \{\langle V_1\rangle_M/\langle V_2\rangle_M\}_{(Z_0,0)} = -f_M(Z_0);$$

and (b) if $V_2(\mathbf{x})>0$, then

$$\min\{\langle V_1\rangle_M/\langle V_2\rangle_M\}_{(Z_1,f_M(Z_1))} = \{\langle V_1\rangle_M/\langle V_2\rangle_M\}_{(Z_0,0)} = -f_M(Z_0).$$

Lemma 1 is one of the results we need to be able to derive approximations to the eigenvalues from the integral representation to the CET's [Eqs. (11) and (12)]. Our strategy can be stated as follows: suppose, without loss of generality, that the problem of $H(Z_1,0)$ can be solved exactly, i.e., all its eigenvalues, as well as all expectation values involving its eigenfunctions can be evaluated in a closed form. Accordingly, even though Eq. (11) cannot be solved exactly, a bound can be given for the integral by using the expectation values computed with the eigenfunctions of $H(Z_0,0)$. Lemma 1 gives us some bounds for the integrand according to the potential. However, to derive approximations to the eigenvalues we need a further result.

Using the VT and the normalization of the eigenfunctions of $H(Z_1,0)$ for all Z_1, it is simple to prove the following scaling law for them:

$$\psi_m(Z_0,0;\mathbf{x}) = a^{-1/2}\psi_m(Z_0a^{2+N_1},0;\mathbf{x}/a). \tag{18}$$

We use Eq. (18) to prove the following lemma.

Lemma 2:

$$\operatorname{sgn}\left[\frac{\partial\{\langle V_1\rangle_m/\langle V_2\rangle_m\}_{(Z_0,0)}}{\partial Z_0}\right] = \operatorname{sgn}\left[\frac{N_2-N_1}{2+N_1}\right].$$

Proof: Invoking the scaling law (18) and the homogeneity properties of the potential, and choosing the scale factor as $a=Z_0^{-1/(2+N_1)}$, we get

$$\{\langle V_1\rangle_m/\langle V_2\rangle_m\}_{(Z_0,0)} = (Z_0)^{(N_2-N_1)/(2+N_1)}C, \tag{19a}$$

where

$$C=\{\langle V_1\rangle_m/\langle V_2\rangle_m\}_{(1,0)}, \tag{19b}$$

from which the lemma follows.

We are now in the position to use Eq. (11) to provide a first simple analytic upper bound to the eigenvalues. Let us consider, for instance, that both V_1 and V_2 are positive. Here, and in following sections, we confine ourselves to this case, to which the formalism has not been applied up till now. The case of both V_1 and V_2 negative has been the subject of a related analysis (in the particular case of diatomic molecules), even though not following the present formulation.[22,23] The extension to negative potentials is immediate.

According to Lemma 1, from Eq. (11) we deduce $[V_2(\mathbf{x})>0]$

$$E_M(Z_1,Z_2)=E_M(Z_0,0) \Rightarrow Z_2\geqslant(Z_0-Z_1)\{\langle V_1\rangle_M/\langle V_2\rangle_M\}_{(Z_0,0)}. \tag{20}$$

Notice that, even though the right-hand side of the inequality (20) can be computed in principle for any value of Z_0, the actual value of Z_0 is unknown. Our goal in the rest of this section is to approximate $E_M(Z_1,Z_2)$ by using an approximate Z_0.

Let us define a new Z_0^* so that the equality is reached in (20):

$$Z_2=(Z_0^*-Z_1)\{\langle V_1\rangle_M/\langle V_2\rangle_M\}_{(Z_0^*,0)}, \tag{21}$$

that is, the new value Z_0^* satisfies

$$(Z_0^*-Z_1)\{\langle V_1\rangle_M/\langle V_2\rangle_M\}_{(Z_0^*,0)} \geqslant(Z_0-Z_1)\{\langle V_1\rangle_M/\langle V_2\rangle_M\}_{(Z_0,0)}. \tag{22}$$

Let us consider now the class of potentials $V_2(\mathbf{x})$ for which the Hamiltonian $H(0,Z_2)$ has eigenvalues that are monotonously increasing functions of Z_2, i.e., $2+N_2>0$ $[V_2(\mathbf{x})>0]$. In this case, Lemma 2 assures us that if $N_2\geqslant N_1$, then inequality (22) implies $Z_0^*>Z_0$. As a consequence we get

$$U(Z_1,Z_2)=E_M(Z_0^*,0) = (Z_0^*)^{2/(N_1+2)}E_M(1,0)\geqslant E_M(Z_1,Z_2). \tag{23}$$

Equation (23) holds also if $V_2(\mathbf{x})<0$. Our conclusion can be summarized in a theorem.

Theorem 1: Let H be a Hamiltonian satisfying conditions (i)–(vi), as well as the conditions $2+N_2>0$ and $N_2\geqslant N_1$ for the potential. Let M be the quantum number of the lowest state in a given manifold of eigenstates of H with distinct symmetry. Then, the function $E_M(Z_0^*,0)$, with Z_0^* the only real positive root of Eq. (21), is an upper bound to the eigenvalue $E_M(Z_1,Z_2)$ for all $(Z_1,Z_2)\in D$.

This theorem provides the simplest example of a class of approximations to eigenvalues that will be discussed in this paper (see Sec. III).

As discussed previously, we have considered the Hamiltonian $H(Z_0^*,0)$ to be a reference Hamiltonian, in the sense that its Schrödinger equation can be solved exactly. In this context its eigenfunctions $\psi_m(Z_0^*,0;\mathbf{x})$ are used to approximate the eigenvalues of $H(Z_1,Z_2)$, which resembles the SVM method.[10–14] Notice, however, that we have not proceeded variationally: neither is $E_M(Z_0^*,0)$ an expectation value of $H(Z_1Z_2)$, nor is Z_0^* obtained by means of variational optimization.

It would not be surprising to obtain an upper bound to $E_M(Z_1,Z_2)$ if we consider that $\psi_M(Z_0^*,0;\mathbf{x})$ is used somehow as a "trial function." However, in the context of the SVM the approximate energy has the correct analytic behavior (as fixed by the HFT and VT's) only when the parameter is optimized variationally. This is *not* our case; nonetheless, we will show that the CET method leads, indeed, to nonvariational approximations with the appropriate analytic structure.

Theorem 2: The function $E_M(Z_0^*,0)$, with Z_0^* the only real positive solution of Eq. (21), is a solution of the following differential equation:

$$E=\frac{2+N_1}{2}Z_1\left(\frac{\partial E}{\partial Z_1}\right)_{Z_2}+\frac{2+N_2}{2}Z_2\left(\frac{\partial E}{\partial Z_2}\right)_{Z_1}, \tag{24}$$

which is the same equation derived after combining the VT and HFT for the solutions of the Hamiltonian (1).

Proof: We first show the scaling law that Z_0^* satisfies with Z_1 and Z_2. Using (19), we rewrite Eq. (21) as follows:

$$Z_2 = (Z_0^* - Z_1)(Z_0^*)^{(N_2 - N_1)/(2+N_1)}C. \tag{25}$$

Introducing now the scaling $Z_0^* = aZ_0'^*$, and $Z_1 = aZ_1'$ into (25), and choosing the scale factor as $a = (Z_2)^{(2+N_1)/(2+N_2)}$, we get

$$1 = (Z_0'^* - Z_1')(Z_0'^*)^{(N_2 - N_1)/(2+N_1)}C, \tag{26}$$

which shows that Z_0^*, as a function of Z_1 and Z_2, satisfies

$$Z_0^*(Z_1,Z_2) = (Z_2)^{(2+N_1)/(2+N_2)} Z_0^*(Z_1(Z_2)^{-(2+N_1)/(2+N_2)},1). \tag{27}$$

Using Eq. (4a), we deduce from (27) that our upper bound $E_M(Z_0^*,0) = U(Z_1,Z_2)$ satisfies the equality

$$U(Z_1,Z_2) = (Z_2)^{2/(2+N_2)} U(Z_1(Z_2)^{-(2+N_1)/(2+N_2)},1). \tag{28}$$

Equation (28) is the correct scaling law expected for the exact eigenvalue according to the standard Symanzik scaling.[24] From (28) one gets

$$Z_1\left(\frac{\partial U(Z_1,Z_2)}{\partial Z_1}\right)_{Z_2} = Z_1(Z_2)^{-N_1/(2+N_2)}\frac{\partial U(b,1)}{\partial b}; \tag{29a}$$

$$Z_2\left(\frac{\partial U(Z_1,Z_2)}{\partial Z_2}\right)_{Z_1} = \frac{2}{2+N_2}(Z_2)^{2/(2+N_2)}U(b,1) - \frac{2+N_2}{2+N_1}Z_1(Z_2)^{-N_1/(2+N_2)}\frac{\partial U(b,1)}{\partial b}, \tag{29b}$$

where $b = Z_1(Z_2)^{-(2+N_1/(2+N_2)}$. Combining Eqs. (29) we complete the proof.

Theorem 2 establishes the fact that the upper bound possesses the correct analytic behavior with the real parameters contained in the Hamiltonian. In other words, it satisfies "simultaneously" the VT and HFT through a nonvariational approach. It is worth commenting that this result holds for all quantum numbers m, not only M, even though the bound (23) might not be true.

The similarity between $U(Z_1,Z_2)$ and the exact eigenvalues can be demonstrated in a more detailed way if we compute the Taylor expansion of the former in power series of Z_2. If we write the parameter Z_0^* as $Z_0^* = Z_0^{(0)} + Z_0^{(1)}Z_2 + Z_0^{(2)}Z_2^2 + \cdots$, introduce it into (21), and collect the coefficients premultiplying each power of Z_2, we can obtain from Eq. (23) the following expansion:

$$U(Z_1,Z_2) \approx E_m(Z_1,0) + Z_2A_2 + Z_2^2\{(N_1 - N_2)A_2^2/E_m(Z_1,0) - N_1A_2/(2+N_1)\} + O(Z_2^3), \tag{30}$$

where $A_2 = \{\langle V_2\rangle_m\}_{(Z_1,0)}$. Clearly, this expansion is correct up to the first order.

We can compare the above result with the one obtained from the variational method (SVM). To this end we consider a trial function that is an eigenfunction of $H(z,0)$. The function is $\psi_m(z,0;\mathbf{x})$, where the real number z will be obtained variationally. Invoking the VT for $H(z,0)$, a simple calculation shows that the variational functional is given by

$$\epsilon(Z_1,Z_2) = (N_1/2)\{\langle V_1\rangle_m\}_{(1,0)}q^{-2} + Z_1\{\langle V_1\rangle_m\}_{(1,0)}q^{N_1} + Z_2\{\langle V_2\rangle_m\}_{(1,0)}q^{N_2}, \tag{31}$$

where $q = z^{-1/(2+N_1)}$. The parameter q takes its optimum value in the stationary condition

$$\frac{\partial\epsilon(Z_1,Z_2)}{\partial q}(q = q^*) = 0;$$

that is,

$$1 = Z_1(q^*)^{N_1+2} + Z_2(N_2/N_1) \times\{\langle V_1\rangle_m/\langle V_2\rangle_m\}_{(1,0)}(q^*)^{N_2+2}. \tag{32}$$

Substitution of q^* into (31) gives an approximation to the eigenvalues that satisfies both VT and HFT, as is well known. Furthermore, its expansion in a power series of Z_2 leads to the following result [notation the same as in Eq. (30)]:

$$\epsilon(Z_1,Z_2) = E_m(Z_1,0) + Z_2A_2 - Z_2^2N_2^2A_2^2/4N_1E_m(Z_1,0) + O(Z_2^3), \tag{33}$$

which is correct up to first order. Notice that the second-order term is not only incorrect but it also differs from the result (30).

Let us summarize the conclusion from the above analysis: the elementary bound (23), obtained from the properties of the boundaries of the level sets in the parameter space D, differs from the SVM result. Both procedures employ *the same* input information, that is, an eigenfunction of the "reference" Hamiltonian $H(Z_1,0)$. However, in the case of the SVM this function is optimized variationally, whereas in the CET method it is not. Despite this difference, both approximations share the following properties: (1) *all* the eigenvalues are described by approximate expressions satisfying the HFT–VT differential equation, which assures the correct dependence with all the parameters in the Hamiltonian; (2) the approximate expressions reproduce the RSPT in power series of Z_2 up to first order; and (3) both give rigorous upper bounds to the energies of the lowest eigenstate of a manifold with distinct symmetry.

Such profound coincidence between a variational and a simple nonvariational method has not been noticed before, as far as we know.

From the numerical point of view, the bounds derived in this section [Eqs. (21) and (23)] are not better than the variational ones. This is not surprising, since the starting bound (20) is the most elementary one that can be found for the integral representation of the CET function $f_m(Z_1)$. As shown in the next section, other expressions obtained from Eq. (11) allow one to improve the accuracy of the approximation, without losing any of the fundamental properties mentioned above.

III. APPROXIMATE EXPRESSIONS FOR EIGENVALUES: GENERAL ANHARMONIC OSCILLATORS

In this section we discuss the derivation of a family of approximate expressions for eigenvalues of parameter-dependent systems, using the main ideas displayed in the previous section.

For the sake of succinctness we consider here a concrete example: the general anharmonic oscillator. This is a one-dimensional model ($\mathbf{x} = x$), given by

$$V_1(\mathbf{x}) = x^2, \quad V_2(\mathbf{x}) = x^{2n}, \quad n = 1,2,3,\ldots . \tag{34}$$

The potential defined by (34), as well as the spectrum of its corresponding Hamiltonian, satisfies the required properties discussed in the previous section. Of course, the case $n = 1$ is trivial. The case $n \geqslant 2$ has no analytical solution, and it has been the subject of numerous discussions in the context of approximate methods (see, for instance, Refs. 5 and 25–28, and others quoted therein). Moreover, the model is a useful one in molecular spectroscopy (see, for instance, Refs. 29–32) as well as in field theory (for example, Refs. 33 and 34). In our case we are only concerned here with the derivation of extremely simple expressions for the eigenvalues, derived from basic principles assuring them the correct analytical behavior with both Z_1 and Z_2. It is worth reiterating that our aim is *not* to compute eigenvalues with large precision. On the contrary, we want to show how an approach different from the variational method can be useful to fix the overall main analytical structure of the eigenvalues, as functions of the physical parameters defining the system. Furthermore, we are more interested in how the expressions obtained can be improved by introducing complementary information about the system, once the basic qualitative behavior of the eigenvalues is guaranteed.

For the system (1) with the potential (34), the condition defining the CET is

$$E_m(Z_0,0) = E_m(Z_1,f_m(Z_1)) = (2m+1)Z_0^{1/2}, \quad Z_1 \in [0,Z_0). \tag{35}$$

According to Eq. (11), the function f_m is represented by the following integral equation:

$$Z_2 = f_m(Z_1) = \int_{Z_1}^{Z_0} \left\{ \frac{\langle x^2 \rangle_m}{\langle x^{2n} \rangle_m} \right\}_{(s,f_m(s))} ds. \tag{36}$$

In Sec. II we analyzed the case of bounding the integral (36) by its minimum value. To obtain better approximations to it, and consequently to the eigenvalues, we propose the following strategy.

(i) The expectation values in (10) are approximated variationally, using the SVM with a $\psi_m(z,0;x)$ wave function.

(ii) These expectation values are then used as approximations to the integrand in Eq. (36). This is achieved by determining the optimum z^* under a double condition: the variational minimum is satisfied *and* the variational energy is made equal to the value of energy defining the CET $[\epsilon(Z_1,Z_2) = (2m+1)Z_0^{1/2}]$.

(iii) Finally the integral obtained is used to recompute approximations to the eigenvalues $E_m(Z_1,Z_2)$.

Proceeding as in Sec. II, we compute the variational expectation value of a Hamiltonian describing a generic point $(s,f_m(s))$ on the CET [cf. Eq. (31)]:

$$\epsilon(s,f_m(s)) = [(2m+1)/2]z^{1/2} + [(2m+1)/2]s/z^{1/2} + C_{n,m} f_m(s) z^{-n/2}, \tag{37}$$

where $C_{n,m}$ stands for the expectation value $\{\langle x^{2n}\rangle_m\}_{(1,0)}$, computed for the mth eigenstate of $H(1,0)$. The first of these elements are[4]

$$C_{1,m} = (2m+1)/2; \quad C_{2,m} = \tfrac{3}{8}\{1 + (2m+1)^2\}; \text{ etc.} \tag{38}$$

The variational condition gives us the optimum z^* as a real positive root of the equation

$$(z^*)^{1/2} = s/(z^*)^{1/2} + 2nC_{n,m} f_m(s) (z^*)^{-n/2}/(2m+1). \tag{39}$$

On the other hand, the condition for belonging to the CET leads us to the following equality:

$$Z_0^{1/2} = s/(z^*)^{1/2} + (n+1)C_{n,m} f_m(s)(z^*)^{-n/2}/(2m+1). \tag{40}$$

Combining Eqs. (39) and (40) we deduce a relationship between the variationally optimized parameter z^* and the parameters Z_0 and s. This equation (a quadratic equation) must be satisfied in order to have a variational approximation to the energy lying on the required CET. A single root of the equation verifies the correct properties for the function $f_m(s)$, and this gives us

$$(z^*)^{1/2} = [n/(n+1)]Z_0^{1/2} \times \{1 + [1 - (n^2-1)s/n^2 Z_0]^{1/2}\}. \tag{41}$$

On the other hand, using Eqs. (18) and (19) (and the definition of $C_{n,m}$), we can now approximate the integrand in Eq. (36) with the variational wave function:

$$\{\langle x^2\rangle_m / \langle x^{2n}\rangle_m\}_{(s,f_m(s))} \approx \{\langle x^2\rangle_m / \langle x^{2n}\rangle_m\}_{(z^*,0)} = [(2m+1)(z^*)^{(n-1)/2}]/2C_{n,m}, \tag{42}$$

where s, Z_0, and z^* are linked as shown in (41). Introducing (42) into (36), and using (41), we obtain

$$Z_2 \approx \frac{2m+1}{2C_{n,m}} \left[\frac{n}{n+1}\right]^{n-1} Z_0^{(n-1)/2} \times \int_{Z_1}^{Z_0} \left\{ 1 + \left[1 - \frac{(n^2-1)s}{n^2 Z_0} \right]^{1/2} \right\}^{n-1} ds. \tag{43}$$

Equation (43) is our starting point to deduce a family of approximate expressions for the eigenvalues. For a given pair (Z_1,Z_2), Eq. (43) is an equality from which Z_0 can be obtained, and from it an approximation to the eigenvalues.

Of course, the solution of Eq. (43) will lead to the SVM result. That is, in this case the eigenvalues $E_m(Z_1,Z_2)$ are approximated by the function $(2m+1)Z_0^{1/2}$, which coincides with the result of Eqs. (31) and (32). We denote this approximation to the eigenvalues by $U_1(Z_1,Z_2)$.

New expressions for the eigenvalues can be deduced by bounding the SVM result from above and below. These expressions have the same analytic behavior as the SVM eigenvalue, and, as we shall see, may lead to more accurate approximations to the exact result.

(a) Upper bounds: A bound for the integral in Eq. (43) can be obtained as follows:

$$\int_{Z_1}^{Z_0}\left\{1+\left[1-\frac{(n^2-1)s}{n^2Z_0}\right]^{1/2}\right\}^{n-1}ds$$
$$\geqslant\left[\frac{n+1}{n}\right]^{n-1}(Z_0-Z_1). \qquad (44)$$

This result and Eq. (43) imply that the value of Z_0 computed as a root of the equation,

$$Z_2=[(2m+1)/2C_{n,m}]Z_0^{(n-1)/2}(Z_0-Z_1), \qquad (45)$$

will be larger than the corresponding root of Eq. (43) (of course, for the same values of n, m, Z_1, and Z_2). In other words, the approximation to the eigenvalues $U_2(Z_1,Z_2)=(2m+1)Z_0^{1/2}$, with Z_0 obtained from (45), satisfies the property

$$U_2(Z_1,Z_2)\geqslant U_1(Z_1,Z_2). \qquad (46)$$

It is immediate to see that $U_2(Z_1,Z_2)$ is simply the approximation to the eigenvalues we constructed in Sec. II [Eqs. (21) and (23)]. As we know, $U_2(Z_1,Z_2)$ possesses the correct analytic structure. However, it also presents a very appealing feature. In the case of the anharmonic oscillators $U_1(Z_1,Z_2)$ gives only an upper bound to the energy of the ground and first excited states ($m=0,1$) for all Z_1 and Z_2. For other states, $U_1(Z_1,Z_2)$ crosses the exact eigenvalue for some unknown Z_1 and Z_2. On the other hand, we find numerically that our approximation $U_2(Z_1,Z_2)$ is an upper bound for all m.

(b) Lower bounds: The following bound holds for the integral in Eq. (43) for all quantum numbers:

$$\int_{Z_1}^{Z_0}\left\{1+\left[1-\frac{(n^2-1)s}{n^2Z_0}\right]^{1/2}\right\}^{n-1}ds$$
$$\leqslant\left\{1+\left[1-\frac{(n^2-1)Z_1}{n^2Z_0}\right]^{1/2}\right\}^{n-1}(Z_0-Z_1). \qquad (47)$$

This result implies that the value of Z_0 obtained from the equation

$$Z_2=[(2m+1)/2C_{n,m}][n/(n+1)]^{n-1}$$
$$\times Z_0^{(n-1)/2}(Z_0-Z_1)$$
$$\times\{1+[1-(n^2-1)Z_1/n^2Z_0]^{1/2}\}^{n-1} \qquad (48)$$

will be smaller than the one obtained from Eq. (43). That is, the approximation to the eigenvalues $U_3(Z_1,Z_2)=(2m+1)Z_0^{1/2}$, with Z_0 computed from (49), satisfies the following inequality:

$$U_3(Z_1,Z_2)\leqslant U_1(Z_1,Z_2). \qquad (49)$$

A simple analysis of Eq. (48) shows that $U_3(Z_1,Z_2)$ satisfies also the correct scaling law (28), i.e., it possesses the same qualitative dependence with Z_1 and Z_2 as the exact eigenvalues, as well as $U_1(Z_1,Z_2)$ and $U_2(Z_1,Z_2)$. It is clear that there also exists a large set of reasonable approximations to the integral in Eq. (48) that could lead to the same analytical result. For instance, replacing the integrand by its mean value would lead to the same scaling law.

TABLE I. Approximations to the ground state energy of the anharmonic quartic oscillator: $H=p^2+x^2+Z_2x^4$.

Z_2	$U_1(1,Z_2)$[a]	$U_2(1,Z_2)$[b]	$U_3(1,Z_2)$[c]	$U_4(1,Z_2)$[d]	$E_0(1,Z_2)$[e]
10^{-3}	1.000 7489	1.000 7492	1.000 7486	1.000 7487	1.000 7487
10^{-1}	1.066 204	1.067 923	1.064 709	1.065 3856	1.065 2855
1	1.403 323	1.431 127	1.383 407	1.392 7179	1.392 3516
10	2.488 624	2.601 244	2.413 049	2.449 4608	2.449 1741
10^4	23.320 33	24.675 64	22.424 68	22.861 6419	22.861 6089

[a] $Z_0^{1/2}$; Z_0 from Eq. (43) ($m=0$) (SVM approximation).
[b] $Z_0^{1/2}$; Z_0 from Eq. (45) ($m=0$) (see also Sec. II).
[c] $Z_0^{1/2}$; Z_0 from Eq. (48) ($m=0$).
[d] $Z_0^{1/2}$; Z_0 from Eqs. (50) and (51) ($m=0$).
[e] Exact results.[35]

Furthermore, it is easily proved that $U_3(Z_1,Z_2)$ also gives rise to the correct RSPT coefficient up to first order. Moreover, we find numerically that $U_3(Z_1,Z_2)$ satisfies not only (50), but it is also a lower bound to the exact eigenvalues for all quantum numbers and values of the parameters in the Hamiltonian. We have not been able to obtain a rigorous proof for this last numerical observation. The function $U_3(Z_1,Z_2)$ is an extremely simple lower bound; however, its analytical similarity with the exact result makes it attractive and useful.

Tables I and II show, respectively, the numerical results obtained for the ground ($m=0$) and second excited state ($m=2$) of the quartic anharmonic oscillator ($n=2$) in a wide range of Z_2 values ($Z_1=1$). We have chosen $m=0$ and $m=2$ to compare two very different situations. In the first case the SVM gives an upper bound for the ground state energy for all Z_2, but it crosses the exact result for some Z_2 in the case of $m=2$.

For the ground state (Table I) we notice that the lower bound $U_3(Z_1,Z_2)$ is closer to the exact result[35] than the variational result. Although the upper bound is distant, it shows an interesting feature: both upper and lower bounds have an almost constant relative deviation from the exact result for Z_2 not too small. As shown below, one can take advantage of this fact.

For the second excited state (Table II) the variational result shows a fortuitous good agreement with the exact result, even though the deviation from it is not uniform. Once

TABLE II. Approximations to the energy of second excited state of the anharmonic quartic oscillator: $H=p^2+x^2+Z_2x^4$.

Z_2	$U_1(1,Z_2)$[a]	$U_2(1,Z_2)$[b]	$U_3(1,Z_2)$[c]	$U_4(1,Z_2)$[d]	$E_2(1,Z_2)$[e]
10^{-3}	5.009 7123	5.009 722	5.009 703	5.009 7117	5.009 7119
10^{-1}	5.748 005	5.781 986	5.721 008	5.747 6354	5.747 9593
1	8.647 038	8.923 591	8.456 671	8.654 2276	8.655 0500
10	16.602 308	17.447 390	16.039 486	16.635 3495	16.635 9215
10^4	160.283 207	169.609 719	154.120 230	160.685 8479	160.685 9126

[a] $5Z_0^{1/2}$; Z_0 from Eq. (43) ($m=2$) (SVM approximation).
[b] $5Z_0^{1/2}$; Z_0 from Eq. (45) ($m=2$) (see also Sec. II).
[c] $5Z_0^{1/2}$; Z_0 from Eq. (48) ($m=2$).
[d] $5Z_0^{1/2}$; Z_0 from Eqs. (50) and (51) ($m=2$).
[e] Exact results.[35]

again, the lower bound is closer to the exact eigenvalues than the upper bound. It is noteworthy that the relative deviation tends to be again somewhat systematic.

The above numerical results suggest that one can use the approximate expressions $U_i(Z_1,Z_2)$, $i = 1,2,3$, as functional representations for the eigenvalues in which one could introduce information provided by some other sources. This has been done previously within a context similar to the SVM.[4,5,36,37] Here we illustrate briefly how this can be accomplished considering the function $U_3(Z_1,Z_2)$, for which the numerical results were better, with the limit of the eigenvalues in the purely anharmonic regime.

Let us suppose that the value of $E_m(0,1)$ is known for an eigenstate of an anharmonic oscillator. We can slightly modify Eq. (48) in order to compute a new value of Z_0, so that the approximation to the eigenvalues obtained from it satisfies the following properties: (i) the scaling law is not affected; (ii) the first-order RSPT coefficient is predicted correctly (this coefficient coincides with $C_{n,m}$ [cf. Eq. (38)]) and (iii) the correct limit of the eigenvalues in the purely anharmonic regime is also predicted properly [in our case, this implies that

$$(Z_2)^{-2/(2+N_2)}U_3(0,Z_2) = E_m(0,1)].$$

A simple way to reach the above goals is to rewrite Eq. (49) as follows:

$$\begin{aligned} k'Z_2 = {} & [(2m+1)/2C_{n,m}][n/(n+1)]^{n-1} \\ & \times Z_0^{(n-1)/2}(Z_0 - Z_1) \\ & \times\{1 + k[1-(n^2-1)Z_1/n^2Z_0]^{1/2}\}^{n-1}. \end{aligned} \tag{50}$$

The function Z_0 computed from (50) satisfies the expected scaling law with respect to Z_1 and Z_2. This fact guarantees that the first requirement above is met. The constants k and k' must then be determined so that the other two conditions are satisfied. A simple computation shows that

$$k = (P_{n,m} - n)/(1 - P_{n,m}), \tag{51a}$$

$$k' = Z_1^{n/2}(n+1)^{1-n}[(P_{n,m} - nP_{n,m})/(1-P_{n,m})]^{n-1}, \tag{51b}$$

where

$$\begin{aligned} P_{n,m} = {} & nE_m(0,1)^{(n+1)/(n-1)}Z_1^{-n/2(n-1)} \\ & \times[2C_{n,m}(2m+1)^n]^{1/(1-n)}. \end{aligned}$$

After solving Eq. (50) for Z_0, we find a new approximation to the eigenvalues, $U_4(Z_1,Z_2) = (2m+1)Z_0^{1/2}$.

We used the results in Ref. 38 for $E_0(0,1)$ and $E_2(0,1)$ to compute $U_4(Z_1,Z_2)$ for the quartic anharmonic oscillator. The corresponding numerical values appear in the fifth column of Tables I and II. It is worth noticing that, even though we have only introduced as a new feature the correct correlation between the spectra of $H(0,Z_2)$ and $H(Z_1,0)$, the accuracy of the approximation is remarkable. In fact, as far as we know, it is one of the simplest and most accurate expressions available for the eigenvalues of the anharmonic oscillators.

These results are a clear illustration of the way in which the CET method can be employed to provide new simple analytical approximations to eigenvalues. It shows that the variational condition over a trial function is unnecessary to satisfy the HFT–VT differential equation, and the RSPT series up to first order, as long as the function is used to represent approximately the constant energy trajectory.

It is worth commenting that other nonvariational techniques (such as the JWKB semiclassical method) may lead to approximations to the eigenvalues also satisfying an equation similar to (24).[17] However, unlike the CET approach, they do not mimic so well the exact eigenvalue, because they fail at predicting (from construction) the correct coefficient at first-order RSPT.

IV. FURTHER COMMENTS AND CONCLUSIONS

In this section we shall discuss briefly some problems connected to the extension of the method. Some of the properties initially required for the potential in Sec. II can be relaxed to apply the CET method in a more general framework. However, there are some very strong conditions that cannot be removed if one is to apply the method in this present formulation. These conditions make it necessary to introduce appropriate modifications in the procedure in order to apply it to some interesting systems.

Let us suppose for instance that the Hamiltonian of interest is given by

$$H(Z_1,Z_2) = -\Delta + Z_1V_1(\mathbf{x}) + Z_2V_2(\mathbf{x}) + V_3(\mathbf{x}), \tag{52}$$

where $V_3(\mathbf{x})$ is some function independent from Z_1 and Z_2 (positive real parameters). Suppose that the Hamiltonians $H(Z_1,0)$ and $H(0,Z_2)$ have discrete spectra for all $Z_1 \in D_1$ and $Z_2 \in D_2$, respectively. In this case condition (iii) (Sec. II) for the potential is no longer a necessary condition to connect both spectra. However, it can be replaced by the following one:

$$\begin{aligned} & \min_{Z_2 \in D_2} \min_{\{\mathbf{x}\}}[Z_2V_2(\mathbf{x}) + V_3(\mathbf{x})] \\ & \quad < \max_{Z_1 \in D_1} \max_{\{\mathbf{x}\}}[Z_1V_1(\mathbf{x}) + V_3(\mathbf{x})]. \end{aligned} \tag{53}$$

Inequality (53) is a necessary condition, even though a not very strong one. With this small extension the CET method can be applied paralleling the derivation in previous sections. The model (52) includes some systems of interest as the electronic Hamiltonian (in the Born–Oppenheimer approximation) for many-electron diatomic molecules. In this case, $V_3(\mathbf{x})$ represents the electron–electron repulsion term. These problems have been discussed in a more restricted framework in Refs. 22 and 23.

Another case of interest is given by the family of Hamiltonians containing two different classes of parameters. For instance, consider the case of $H(Z_1,Z_2)$ given by

$$H(Z_1,Z_2) = -\Delta + Z_1V_1(\mathbf{x}) + Z_2V_2(\mathbf{x};\mathbf{r}) + V_3(\mathbf{x}), \tag{54}$$

containing a linear dependence on Z_1 and Z_2, and a nonlinear dependence in a set of parameters $\mathbf{r}$. For these problems it is sometimes more interesting to describe the behavior of the eigenvalues as functions of $\mathbf{r}$, for fixed values of Z_1 and Z_2. The constant energy trajectories should be here studied as functions of $\mathbf{r}$; for a given pair (Z_1,Z_2) there can exist, in general, infinitely many CET's connecting the spectra of $H(Z_1,0)$ and $H(0,Z_2)$. Some results concerning the properties of these particular CET's are present in previous references.[21–23]

We notice here that there exist certain systems depending on two parameters, as in (1), for which the CET method cannot be applied in the present formulation. For example, if condition (iii) (Sec. II) is not fulfilled [i.e., $\mathrm{sgn}(V_1(\mathbf{x})) \neq \mathrm{sgn}(V_2(\mathbf{x}))$], then the two limiting spectra cannot be joined by a CET. Examples of problems with these characteristics include the Zeeman effect on hydrogen (with a uniform static magnetic field)[39] and the "quarkonium" model potential.[40–42] In the first case the relevant part of the first Hamiltonian is given by

$$V_1(\mathbf{x}) = -1/\|\mathbf{x}\|, \quad V_2(\mathbf{x}) = (x_1^2 + x_2^2),$$
$$\mathbf{x} = (x_1,x_2,x_3), \tag{55}$$

when the field is aligned in the x_3 direction, whereas in the last case we have the same $V_1(\mathbf{x})$, but

$$V_2(\mathbf{x}) = \|\mathbf{x}\|^n, \quad n \geqslant 0.$$

Nevertheless, it is still possible to define CET's for these systems, if the condition to have a bounded curve [$f(Z_1) \leqslant Z_0'$, for $Z_1 \leqslant Z_0$] is removed. In this case, Eqs. (6) will no longer be valid. These generalized unbounded CET's can be used again to provide some elementary bounds for the eigenvalues. This problem will be discussed elsewhere in a forthcoming paper.

ACKNOWLEDGMENTS

We would like to thank Dr. Francisco M. Fernández for sending us a copy of the manuscript of Ref. 9 prior to publication, and for an interesting and helpful correspondence.

This work was supported by a research grant from the Natural Sciences and Engineering Research Council (NSERC) of Canada.

[1] G. Rosen, Phys. Rev. A **20**, 1287 (1979).
[2] H. Orland, Phys. Rev. Lett. **42**, 285 (1979).
[3] F. M. Fernández and E. A. Castro, Phys. Rev. A **27**, 2735 (1983).
[4] F. M. Fernández, G. A. Arteca, and E. A. Castro, Physica A **122**, 37 (1983); Int. J. Quantum Chem. **25**, 1023 (1984).
[5] G. A. Arteca, F. M. Fernández, and E. A. Castro, J. Math. Phys. **25**, 932 (1984); Z. Phys. A **315**, 255 (1984).
[6] K. Banerjee, Proc. R. Soc. London Ser. A **380**, 489 (1982).
[7] E. Gerck, J. A. C. Gallas, and A. B. d' Oliveira, Phys. Rev. A **26**, 662 (1982).
[8] G. Rosen, Phys. Rev. A **34**, 1556 (1986).
[9] F. M. Fernández and E. A. Castro, Phys. Rev. A **35**, 4861 (1987).
[10] F. M. Fernández and E. A. Castro, Phys. Rev. A **27**, 663 (1983); J. Chem. Phys. **79**, 321 (1983).
[11] V. Fock, Z. Phys. **63**, 855 (1930).
[12] R. McWeeny and C. A. Coulson, Proc. Camb. Philos. Soc. **44**, 413 (1948).
[13] P.-O. Löwdin, J. Mol. Spectrosc. **3**, 46 (1959).
[14] C. C. Gerry an J. Laub, Phys. Rev. A **30**, 1229 (1984).
[15] G. A. Arteca, F. M. Fernández, and E. A. Castro, J. Math. Phys. **25**, 2377 (1984).
[16] G. A. Arteca, F. M. Fernández, and E. A. Castro, J. Phys. A **20**, 2221 (1987).
[17] F. M. Fernández, G. A. Arteca, and E. A. Castro, J. Chem. Phys. **80**, 5659 (1984).
[18] P. G. Mezey, Chem. Phys. Lett. **87**, 277 (1982).
[19] P. G. Mezey, Int. J. Quantum Chem. **22**, 101 (1982); **29**, 85 (1986).
[20] P. G. Mezey, J. Am. Chem. Soc. **107**, 3101 (1985).
[21] G. A. Arteca and P. G. Mezey, J. Chem. Phys. (in press).
[22] G. A. Arteca and P. G. Mezey, Phys. Rev. A **35**, 4044 (1987).
[23] G. A. Arteca and P. G. Mezey, Phys. Lett. A **122**, 483 (1987).
[24] B. Simon, Ann. Phys. (NY) **58**, 76 (1970).
[25] F. T. Hioe, D. MacMillen, and E. Montroll, J. Math. Phys. **17**, 1320 (1976).
[26] B. Hirsbrunner, Helv. Phys. Acta **55**, 295 (1982).
[27] See the issue: Int. J. Quantum Chem. **21** (1) (1982).
[28] A. V. Turbiner, Sov. Phys. Usp. **27**, 668 (1984) [Usp. Fiz. Nauk. **144**, 35 (1984)].
[29] R. P. Bell, Proc. R. Soc. London Ser. A **183**, 328 (1945).
[30] S. I. Chan, J. Zinn, and W. D. Gwinn, J. Chem. Phys. **34**, 1319 (1961).
[31] A. Danti, W. J. Lafferty, and R. C. Lord, J. Chem. Phys. **34**, 294 (1961).
[32] T. Ueda and T. Shimanouchi, J. Chem. Phys. **49**, 470 (1968).
[33] C. M. Bender and T. T. Wu, Phys. Rev. **184**, 1231 (1969).
[34] R. Seznec and J. Zinn-Justin, J. Math. Phys. **20**, 1398 (1979).
[35] K. Banerjee, Proc. R. Soc. London Ser. A **364**, 265 (1978).
[36] P. Pascual, An. Fis. **75**, 77 (1979).
[37] I. K. Dmitrieva and G. I. Plindov, Phys. Lett. A **79**, 47 (1980); Phys. Scr. **22**, 386 (1980).
[38] K. Banerjee, S. P. Bhatnagar, V. Choudhry, and S. S. Kanwal, Proc. R. Soc. London Ser. A **360**, 575 (1978).
[39] R. H. Garstang, Rep. Prog. Phys. **40**, 105 (1977).
[40] E. Eichten, K. Gottfried, T. Kinoshita, K. D. Lane, and F. M. Yan, Phys. Rev. D **17**, 3090 (1978).
[41] C. Quigg and J. L. Rosner, Phys. Rep. **56**, 167 (1979).
[42] J. Killingbeck, J. Phys. A **14**, 1005 (1981).

Helicity eigenstates of a relativistic spin-0 and spin-$\frac{1}{2}$ constituent bound by minimal electrodynamics: Zero orbital angular momentum, zero four-momentum solutions

G. Bruce Mainland
Department of Physics, Ohio State University, Columbus, Ohio 43210

(Received 13 February 1987; accepted for publication 9 September 1987)

Zero four-momentum, helicity eigenstates of the Bethe–Salpeter equation are found for a composite system consisting of a charged, spin-0 constituent and a charged, spin-$\frac{1}{2}$ constituent bound by minimal electrodynamics. The form of the Bethe–Salpeter equation used to describe the bound state includes the contributions from both single photon exchange (ladder approximation) and the "seagull" diagram. Attention is restricted to zero orbital angular momentum states since these appear to be the most interesting physically.

I. INTRODUCTION

Experimentally the electron-muon mass ratio is numerically approximately two-thirds of the electromagnetic fine structure constant. If this is not a coincidence, and charged leptons are composite, electromagnetism must play an important role in binding the constituents. We are therefore motivated to consider two or possibly three constituents bound by electrodynamics.

As a first step toward determining the consequences of such a model, a bound state consisting of a charged particle orbiting a stationary (infinitely massive) charged, magnetic dipole was studied in two space dimensions using the Schrödinger equation.[1] Three encourgaging results were obtained: (1) A bound state with orbital angular momentum $l = 0$ can occur that has a radius smaller than the present experimental limit for the electron. (2) No strongly bound (low mass) states occur except for orbital angular momentum $l = 0$. Thus if one of the constitutents has spin-0 and the other has spin-$\frac{1}{2}$, all low-lying bound states would have spin-$\frac{1}{2}$ as required by the charged lepton mass spectrum. But since an electromagnetic transition between two $l = 0$ states is a forbidden transition,[2] a third encouraging result follows. (3) The model provides a natural explanation for the absence of the decay $\mu \rightarrow e + \gamma$. Because of the results of this preliminary calculation, attention was restricted to a two-body model of charged leptons consisting of a charged, spin-0 boson and a charged, spin-$\frac{1}{2}$ fermion interacting electromagnetically.

A second preliminary calculation[3] was performed using the Klein–Gordon equation to describe the constituent boson moving in two space dimensions under the influence of the electromagnetic field of an infinitely massive fermion. Although the "sizes" of bound states were not determined, the absence of strongly bound states except for $l = 0$ remained. An important new result was the qualitative behavior of the energy spectrum. (4) The energy gaps between successively higher bound states can increase in size. This very unusual bound-state pattern is qualitatively similar to the charged lepton mass spectrum. However, because the Klein–Gordon equation was used, the rate of increase between the energies of successive bound states is larger than that of the observed mass spectrum. On the other hand, had the Dirac equation been employed, the gaps between successive bound-state energy levels would have decreased. By treating the spin-0 and spin-$\frac{1}{2}$ constituents symmetrically in a fully relativistic calculation, it may be possible to obtain energy gaps that agree with the experimental mass spectrum of charged leptons.

If the muon and tau are excited states of the electron, then the constituents must be described relativistically. Only then would it be possible to obtain mass (energy) gaps between successive states that are large compared with the mass (energy) of the most tightly bound state. Also, the small upper limit on the electron's "radius" suggests relativistic binding.

Both of the preliminary calculations just described were performed in two space dimensions in order that the equations could be separated. Also, the first calculation was nonrelativistic while the second was only partially relativistic, so none of the results obtained are rigorously established in four-dimensional space-time and are only suggestive. Nevertheless, the preliminary results indicate it might be worthwhile to consider a charged, spin-0 and a charged, spin-$\frac{1}{2}$ constituent interacting relativistically via minimal electrodynamics using a relativistic equation such as the Bethe–Salpeter equation.

Constructing the "exact" Bethe–Salpeter equation requires considering all Feynman diagrams and is, of course, impossible from a practical standpoint. Usually only single exchange of the binding quanta is considered (ladder approximation). If the contributions of higher-order diagrams to the bound states are small, then the ladder approximation is acceptable physically. Two additional mathematical approximations, which are not necessarily justified physically, are commonly made to make the equation easier to solve analytically. (1) The masses of the constituents are assumed to be equal. (2) The four-momentum of the bound state is taken to equal zero.

Making the above two mathematically motivated approximations, the Bethe–Salpeter equation was solved in the ladder approximation for bound states of a minimally interacting charged, spin-0 constituent and a charged, spin-$\frac{1}{2}$ constituent.[4] Since the four-momentum is zero, the Bethe–Salpeter equation becomes an eigenvalue equation for the

0022-2488/88/010128-11$02.50

coupling constant instead of energy, and the spectrum of the coupling constant was found to be discrete. The solution was obtained by analytically continuing Minkowski space into Euclidean space,[5] projecting four-dimensional Euclidean space onto the surface of a five-dimensional sphere,[6-8] assuming the solution is an infinite sum of hyperspherical harmonics, and integrating the resultant integral using Hecke's theorem.[9]

For the system being considered, the ladder approximation may be unsatisfactory for very tightly bound states. To understand intuitively why, recall that when the Klein–Gordon equation is used to describe the minimal electromagnetic interaction of a spin-0 constituent, the following substitution is made: $(i\partial^\mu)^2 \to (i\partial^\mu - qA^\mu)^2$. In the Bethe–Salpeter equation, the above linear term in A^μ corresponds to single photon exchange (ladder approximation) and the quadratic term $A_\mu A^\mu$ corresponds to the seagull diagram. In the presence of a charged, magnetic dipole, for partially relativistic calculations A^0 is the Coulomb potential and is proportional to $1/r$ while $\mathbf{A}$ is the magnetic potential resulting from the dipole and is proportional to $1/r^2$. Therefore at small distances the seagull term makes a contribution proportional to $1/r^4$ and is especially important. If the radius of the most tightly bound states is sufficiently small then, at least in the partially relativistic calculation, the seagull term cannot be neglected.

There is a second indication that the seagull term must be included. When the Klein–Gordon equation was used to describe the constituent boson moving in two space dimensions under the influence of the electromagnetic field of an infinitely massive fermion,[3] if the quadratic term $A_\mu A^\mu$ had been neglected, the energy gaps between successively higher bound states would have decreased. The quadratic term $A_\mu A^\mu$ is responsible for the unusual energy spectrum in which the energy gaps increase between successively higher-bound states. If solutions to the Bethe–Salpeter equation with this unusual energy spectrum exist, it is likely that the seagull term is responsible.

From the partially relativistic calculation, we are thus motivated to include the seagull contribution to the Bethe–Salpeter equation, but not necessarily contributions from other higher-order diagrams. Because of the behavior of the seagull contribution at small distances, it can be important for tightly bound states even though it is a second-order diagram. However it is unlikely that other higher-order diagrams would make such a significant contribution.

If charged leptons are composite, the most likely candidates are $l = 0$ bound states. In addition to providing a natural explanation for the absence of the decay $\mu \to e + \gamma$, $l = 0$ states are generally more tightly bound. Recall that when the Schrödinger equation is separated in the presence of a spherically symmetric potential, in the radial equation there is an effective repulsive potential proportional to $l(l+1)/r^2$ caused by an effective centrifugal force. When $l = 0$, the effective potential and effective centrifugal force vanish, and the bound state is more tightly bound. This same general effect is to be expected in relativisitic calculations.

For isolated systems, total angular momentum is a good quantum number, but orbital angular momentum usually is not. For example, when the Dirac equation is used to describe an electron with total angular momentum j bound by a Coulomb potential, two values of orbital angular momentum contribute to each solution, $l_1 = j + \frac{1}{2}$ and $l_2 = j - \frac{1}{2}$. For many types of interactions, no bound orbital angular momentum eignestates with $l = 0$ exist. However, because of the specific form of the minimal electrodynamics interaction, if bound states with zero four-momentum are assumed to exist that are eigenstates of orbital angular momentum with $l = 0$, it is possible to separate the equation as shown in Sec. III. But this separation is possible only if the constituent fermion is massless.

No massless, charged fermions exist as free entities, so if the constituent fermion is required to be massless in a finite-energy (realistic) calculation, it must either be bound very tightly or be confined. If, in addition to electrodynamics, a confining force(s) exists, it could play a negligible role in the energy (mass) spectrum of the bound states provided the confining force(s) is negligible in comparison to the electrodynamic forces at distances comparable to the "radii" of the bound states.

If the constituent fermion is massless, the Bethe–Salpeter equation has the property that each term anticommutes with γ_5. Multiplying by the helicity projection operator $\frac{1}{2}(1-\gamma_5)$ or $\frac{1}{2}(1+\gamma_5)$, the Bethe–Salpeter equation becomes a two-component equation for left-handed or right-handed helicity eigenstates, respectively. For $l = 0$, the two-component equations are solved in the zero four-momentum limit. In Sec. IV the ladder approximation is employed and four left-handed helicity eigenstates are found. In Sec. V the seagull contribution is included and five left-handed helicity eigenstates are found. For every left-handed solution there is, of course, a corresponding right-handed solution. In the zero four-momentum limit, the Bethe–Salpeter equation becomes an eigenvalue equation for the coupling constant and, for all these solutions, the eigenvalue spectrum of the coupling constant is continuous.

It is certainly speculative to view the charged leptons as being composite. But if the charged leptons are composite, the neutrinos may also be. It is remarkable that when the constituent fermion is massless, the four-component Bethe–Salpeter equation can be split into a two-component equation for left-handed helicity eigenstates and a two-component equation for right-handed helicity eigenstates. (For most interactions each two-component equation would involve both left- and right-handed helicity eigenstates.)

If a neutrino is both massless and composite, solving the bound-state equation determines the eigenvalue spectrum of some combination of the constituent mass(es) and coupling constant(s). If the eigenvalue spectrum is discrete, then a constraint would exist among the constituent masses and coupling constants. However, if the eigenvalue spectrum is continuous, as it is for the zero four-momentum solutions found here, then no such constraint would exist.

II. BETHE–SALPETER EQUATION INCLUDING BOTH SINGLE PHOTON EXCHANGE AND THE SEAGULL CONTRIBUTION

We consider a spin-0 field $\phi(x)$ that describes a particle with charge Q and mass M, interacting via minimal electrodynamics with a spin-$\frac{1}{2}$ field $\psi(x)$ that describes a particle with charge q and mass m. The (renormalizable) Lagrangian is

$$L = :[(i\partial^\mu - QA^\mu)\phi][(-i\partial_\mu - QA_\mu)\phi^\dagger] - M^2\phi^\dagger\phi + \bar{\psi}\gamma_\mu(i\partial^\mu - qA^\mu)\psi - m\bar{\psi}\psi - \tfrac{1}{4}F_{\mu\nu}F^{\mu\nu}:, \tag{2.1}$$

where $F_{\mu\upsilon} = \partial_\nu A_\mu - \partial_\mu A_\nu$.

The two-particle, Bethe–Salpeter wave function $\chi_{K,\alpha}(x_1,x_2)$ is defined by

$$\chi_{K,\alpha}(x_1,x_2) = \langle 0|T\psi(x_1)\phi(x_2)|K,\alpha\rangle. \tag{2.2}$$

In Eq. (2.2) the symbol T represents time ordering, the symbol K labels the four-momentum K_μ of the bound state, and α labels any other quantum numbers necessary to specify the state. The relative coordinates x^μ are defined by

$$x^\mu = x_1^\mu - x_2^\mu, \tag{2.3}$$

and the center-of-mass coordinates X^μ by

$$X^\mu = \xi x_1^\mu + (1-\xi)x_2^\mu, \tag{2.4}$$

where ξ is a constant. The dependence of $\chi_{K,\alpha}(x_1,x_2)$ on the center-of-mass coordinates factors with the result that $\chi_{K,\alpha}(x_1,x_2)$ can be written as

$$\chi_{K,\alpha}(x_1,x_2) = (2\pi)^{-3/2}e^{-iX^\mu K_\mu}\chi_{K,\alpha}(x), \tag{2.5}$$

where $\chi_{K,\alpha}(x)$ is given by

$$\chi_{K,\alpha}(x) = (2\pi)^{3/2}\langle 0|T\psi[(1-\xi)x]\phi(-\xi x)|K,\alpha\rangle. \tag{2.6}$$

Following standard procedures[10] and including the contributions from both single photon exchange (ladder approximation) and the seagull diagram, the Bethe–Salpeter equation is[11]

$$\begin{aligned}(\gamma^\mu p_\mu + \xi\gamma^\mu K_\mu - m)\{&-[p^\nu - (1-\xi)K^\nu]\\ &\times[p_\nu - (1-\xi)K_\nu] + M^2\}\chi_{K,\alpha}(p)\\ = -iqQ&\int_{-\infty}^{\infty}\frac{d^4q}{(2\pi)^4}\frac{1}{(p-q)^2+i\epsilon}\\ &\times\gamma^\mu[p_\mu + q_\mu - 2(1-\xi)K_\mu]\chi_{K,\alpha}(q) - 4(Qq)^2\\ &\times\int_{-\infty}^{\infty}\frac{d^4q}{(2\pi)^4}\frac{2m-\gamma^\mu q_\mu}{q^2-m^2+i\epsilon}\frac{1}{(p+\xi K-q)^2+i\epsilon}\\ &\times\int_{-\infty}^{\infty}\frac{d^4k}{(2\pi)^4}\frac{1}{(k+\xi K-q)^2+i\epsilon}\chi_{K,\alpha}(k).\end{aligned} \tag{2.7}$$

In the above equation, $\chi_{K,\alpha}(p)$ is the Fourier transform of $\chi_{K,\alpha}(x)$,

$$\chi_{K,\alpha}(p) = \frac{1}{(2\pi)^2}\int_{-\infty}^{\infty} d^4x\, e^{ip\cdot x}\chi_{K,\alpha}(x). \tag{2.8}$$

The terms proportional to qQ and $(qQ)^2$ in (2.7) are, respectively, the contributions from single photon exchange and the seagull diagram.

We now set $K_\mu = 0$ thereby restricting attention to the zero four-momentum limit. To put (2.7) into a form that is more readily solved, we use the analytic properties of the bound-state wave function χ and analytically continue the equation into four-dimensional Euclidean space.[5] We begin by recalling that in the term from single photon exchange, it is possible to rotate the q_0 path of integration 90° counterclockwise in the complex q_0 plane to the straight line from $-i\infty$ to $i\infty$ if p_0 is first rotated 90° counterclockwise in the complex p_0 plane by making the substitution $p_0 \to ip_0$.[5] To verify that it is possible to analytically continue the seagull contribution into Euclidean space, we first note that the singularities in the integrand of the dk_0 integral are identical to those in the single photon exchange term if we make the substitutions $p \to q$, $q \to k$ in the latter term. Thus in the contribution from the seagull diagram, it is possible to rotate the k_0 path of integration to the straight line from $-i\infty$ to $i\infty$ if q_0 can be rotated 90° counterclockwise in the complex q_0 plane. But in the seagull contribution, the singularities in the integrand of the dq_0 integral are the same as these in the single photon exchange term except for two additional poles from the term $(q^2-m^2)^{-1}$. As can be easily checked, the two new poles do not interfere with the rotation of the q_0 path of integration. Since the integral over d^4k yields an analytic function in q, in the seagull contribution it is possible to rotate the q_0 path of integration to the straight line from $-i\infty$ to $i\infty$ if p_0 is first rotated 90° counterclockwise in the complex p_0 plane by making the substitution $p_0 \to ip_0$. Rotating p_0 and changing the paths of integration as indicated above, in the zero four-momentum limit, the Bethe–Salpeter equation takes the following Euclidean form:

$$\begin{aligned}(\tilde{\gamma}\cdot p + m)(p\cdot p + M^2)\chi_\alpha(p)\\ = -\frac{qQ}{(2\pi)^4}&\int_{-\infty}^{\infty} d^4q\frac{1}{(p-q)\cdot(p-q)}\\ &\times[\tilde{\gamma}\cdot(p+q)]\chi_\alpha(q)\\ &+\frac{4q^2Q^2}{(2\pi)^8}\int_{-\infty}^{\infty} d^4q\frac{2m+\tilde{\gamma}\cdot q}{q\cdot q+m^2}\frac{1}{(p-q)\cdot(p-q)}\\ &\times\int_{-\infty}^{\infty} d^4k\frac{1}{(q-k)\cdot(q-k)}\chi_\alpha(k).\end{aligned} \tag{2.9}$$

In the above equation $\chi_\alpha(p) \equiv \chi_{K=0,\alpha}(ip_0,\mathbf{p})$, the Euclidean scalar product $p\cdot p \equiv p^0p^0 + \mathbf{p}\cdot\mathbf{p}$, and $\tilde{\gamma}\cdot p \equiv \tilde{\gamma}^0p^0 + \gamma^ip^i$. The matrix $\tilde{\gamma}^0 \equiv -i\gamma^0$ where the matrices γ^μ are those of Ref. 11.

III. SEPARABILITY OF THE BETHE-SALPETER EQUATION IN THE LADDER APPROXIMATION ASSUMING THE EXISTENCE OF ZERO ORBITAL ANGULAR MOMENTUM, ZERO FOUR-MOMENTUM EIGENSOLUTIONS

In this section we assume the existence of zero orbital angular momentum, zero four-momentum eigensolutions, and then determine the conditions under which the Bethe–Salpeter equation separates. We find that separation occurs only if the spin-$\frac{1}{2}$ constituent is massless. Furthermore, even when the constituent fermion is massless, of all *a priori* possible $l = 0$ eigenstates, only a subset separate. Here, for simplicity, we consider the Bethe–Salpeter equation in the ladder approximation. In Sec. V we include the seagull term and verify that the equation also separates there when the wave function is assumed to have the form determined here.

Since we will eventually set the constituent fermion mass $m = 0$, it is convenient to write the four-component equation (2.9) as two, two-component equations. To accomplish this we write

$$\chi_a \equiv \begin{pmatrix} \chi_u \\ \chi_d \end{pmatrix}, \tag{3.1}$$

define

$$\chi^{(\pm)} = \chi_u \pm \chi_d, \tag{3.2}$$

and note that

$$(1 \pm \gamma_5)\chi \equiv (1 \pm \gamma_5) \begin{pmatrix} \chi_u \\ \chi_d \end{pmatrix} = \begin{pmatrix} \chi^{(\pm)} \\ \pm\chi^{(\pm)} \end{pmatrix}. \tag{3.3}$$

Multiplying (2.9) by $(1 \mp \gamma_5)$ and recalling that γ_5 anticommutes with $\tilde{\gamma}^\mu$,

$$\begin{aligned} &\tilde{\gamma}\cdot p(p\cdot p + M^2) \begin{pmatrix} \chi^{(\pm)}(p) \\ \pm\chi^{(\pm)}(p) \end{pmatrix} \\ &\quad + m(p\cdot p + M^2) \begin{pmatrix} \chi^{(\mp)}(p) \\ \mp\chi^{(\mp)}(p) \end{pmatrix} \\ &\quad = -\frac{qQ}{(2\pi)^4} \int_{-\infty}^{\infty} d^4q \, \frac{\tilde{\gamma}\cdot(p+q)}{(p-q)\cdot(p-q)} \\ &\qquad \times \begin{pmatrix} \chi^{(\pm)}(q) \\ \pm\chi^{(\pm)}(q) \end{pmatrix}. \end{aligned} \tag{3.4}$$

As mentioned previously, for simplicity the contribution from the seagull term is being omitted. Writing the four-component gamma matrices in terms of the Pauli sigma matrices σ^i, (3.4) can be rewritten as the two-component equation

$$\begin{aligned} &(-ip^0 \pm \boldsymbol{\sigma}\cdot\mathbf{p})(p\cdot p + M^2)\chi^{(\pm)}(p) \\ &\quad + m(p\cdot p + M^2)\chi^{(\mp)}(p) \\ &\quad = -\frac{qQ}{(2\pi)^4} \int_{-\infty}^{\infty} d^4q \, \frac{1}{(p-q)\cdot(p-q)} \\ &\qquad \times [-i(p^0+q^0) \pm \boldsymbol{\sigma}\cdot(\mathbf{p}+\mathbf{q})]\chi^{(\pm)}(q). \end{aligned} \tag{3.5}$$

To proceed we introduce polar coordinates

$$\begin{aligned} &p^0 = |p|\cos\theta_2, \quad p^3 = |p|\sin\theta_2\cos\theta_3, \\ &p^1 = |p|\sin\theta_2\sin\theta_3\cos\phi, \quad p^2 = |p|\sin\theta_2\sin\theta_3\sin\phi, \end{aligned} \tag{3.6}$$

with corresponding expressions for the components of q^μ in terms of primed angles. Then

$$\begin{aligned} d^4q &= |q|^3 \sin^2\theta'_2 \sin\theta'_3 \, d|q| d\theta'_2 \, d\theta'_3 \, d\phi' \\ &\equiv |q|^3 d|q| d\Omega'_{(3)}. \end{aligned} \tag{3.7}$$

Defining the four-dimensional unit vectors by

$$\begin{aligned} \hat{u}_{(4)} \equiv (&\cos\theta_2, \ \sin\theta_2\cos\theta_3, \ \sin\theta_2\sin\theta_3\cos\phi, \\ &\sin\theta_2\sin\theta_3\sin\phi), \end{aligned} \tag{3.8a}$$

$$\begin{aligned} \hat{v}_{(4)} = (&\cos\theta'_2 \ \sin\theta'_2\cos\theta'_3, \ \sin\theta'_2\sin\theta'_3\cos\phi', \\ &\sin\theta'_2\sin\theta'_3\sin\phi'), \end{aligned} \tag{3.8b}$$

we find

$$(p-q)\cdot(p-q) = |p|^2 + |q|^2 - 2|p| \, |q|\cos\Theta_{(4)}, \tag{3.8c}$$

where $\Theta_{(4)}$ is the angle between $\hat{u}_{(4)}$ and $\hat{v}_{(4)}$.

The possible angular dependence of the four-component spinor χ_α is known.[4] From the general results [see Ref. 4, Eq. (3.14)], it follows immediately that the most general form of a two-component spinor $\chi_\pm$ that is an eigenstate of orbital angular momentum $l = 0$ is

$$\begin{aligned} \chi^{(\pm)}(p) &= 2f^{(\pm)}(|p|)P^{(2)}_{n,0}(\theta_2)\phi^{(+)}_{j=1/2,m}(\theta_3,\phi), \\ &\qquad n = 0,1,2,\ldots. \end{aligned} \tag{3.9}$$

In the above equation $f^{(\pm)}(|p|)$ is an undetermined function of $|p|$, the factor of 2 is arbitrarily included for later convenience, and $P^{(2)}_{p,r}$ is defined in Ref. 4 [Eq. (A8)]. The two-component spinors $\phi^{(\pm)}_{j,m}$ have the indicated eigenvalues and are given below,

$$\begin{aligned} \phi^{(+)}_{j,m}(\theta_3,\phi) &= \begin{bmatrix} \sqrt{(j+m)/2j} & Y^{m-1/2}_{j-1/2}(\theta_3,\phi) \\ \sqrt{(j-m)/2j} & Y^{m+1/2}_{j-1/2}(\theta_3,\phi) \end{bmatrix}; \\ &\qquad j = l + 1/2; \quad l = 0,1,2,\ldots, \end{aligned} \tag{3.10a}$$

$$\begin{aligned} &\phi^{(-)}_{j,m}(\theta_3,\phi) \\ &\quad = \begin{bmatrix} \sqrt{(j+1-m)/2(j+1)} & Y^{m-1/2}_{j+1/2}(\theta_3,\phi) \\ -\sqrt{(j+1+m)/2(j+1)} & Y^{m+1/2}_{j+1/2}(\theta_3,\phi) \end{bmatrix}; \\ &\quad j = l - 1/2; \quad l = 1,2,\ldots. \end{aligned} \tag{3.10b}$$

The $\phi^{(\pm)}_{j,m}$ can be transformed into each other using the relationship

$$\phi^{(\pm)}_{j,m}(\theta_3,\phi) = (\boldsymbol{\sigma}\cdot\mathbf{p}/|\mathbf{p}|)\phi^{(\mp)}_{j,m}(\theta_3,\phi). \tag{3.11}$$

In (3.10), the $Y^m_l(\theta_3,\phi)$ are spherical harmonics. Using (3.6)–(3.11), (3.5) becomes

$$\begin{aligned} &|p|(|p|^2 + M^2)2f^{(\pm)}(|p|) \\ &\quad \times [-i\cos\theta_2 P^{(2)}_{n,0}(\theta_2)\phi^{(+)}_{1/2,m}(\theta_3,\phi) \pm \sin\theta_2 P^{(2)}_{n,0}\phi^{(-)}_{1/2,m}(\theta_3,\phi)] + m(|p|^2 + M^2)2f^{(\mp)}(|p|)P^{(2)}_{n,0}(\theta_2)\phi^{(+)}_{1/2,m}(\theta_3,\phi) \\ &\quad = -\frac{qQ}{(2\pi)^4}(-ip^0 \pm \boldsymbol{\sigma}\cdot\mathbf{p}) \int \frac{d|q| d\Omega'_{(3)} |q|^3}{|p|^2 + |q|^2 - 2|p| \, |q|\cos\Theta_{(4)}} 2f^{(\pm)}(|q|)P^{(2)}_{n,0}(\theta'_2)\phi^{(+)}_{1/2,m}(\theta'_3,\phi') \\ &\qquad - \frac{qQ}{(2\pi)^4} \int \frac{d|q| d\Omega'_{(3)} |q|^4}{|p|^2 + |q|^2 - 2|p| \, |q|\cos\Theta_{(4)}} \\ &\qquad \times 2f^{(\pm)}(|q|)[-i\cos\theta'_2 P^{(2)}_{n,0}(\theta'_2)\phi^{(+)}_{1/2,m}(\theta'_3,\phi') \pm \sin\theta'_2 P^{(2)}_{n,0}(\theta'_2)\phi^{(-)}_{1/2,m}(\theta'_3,\phi')]. \end{aligned} \tag{3.12}$$

Now from Ref. 4 [Eq. (A23)], for $n \geqslant 1$,

$$\cos\theta P^{(s)}_{n,0}(\theta) = [1/(2n+s)][(n+1)P^{(s)}_{n+1,0}(\theta) + (n+s-1)P^{(s)}_{n-1,0}(\theta)]. \tag{3.13}$$

To determine the relationship for $n = 0$, we first note from the definition of $P^{(s)}_{p,r}$ that

$$P^{(s)}_{0,0}(\theta) = 1, \quad P^{(s)}_{1,0}(\theta) = s\cos\theta, \tag{3.14}$$

implying

$$\cos\theta\, P^{(s)}_{0,0}(\theta) = (1/s)P^{(s)}_{1,0}(\theta). \tag{3.15}$$

With the convention $P^{(s)}_{p,r} = 0$ if $p < r$, (3.13) includes (3.15) as a special case and is therefore valid for all $n = 0,1,2,\ldots$. From Ref. 4 [Eq. (A24)], for $n \geqslant 2$,

$$\sin\theta\, P^{(s)}_{n,0}(\theta) = [1/(2n+s)][P^{(s)}_{n+1,1}(\theta) - P^{(s)}_{n-1,1}(\theta)]. \tag{3.16}$$

To determine the relationship for $n = 0$ and 1, we use the definition of $P^{(s)}_{p,r}$ and calculate

$$P^{(s)}_{1,1}(\theta) = s\sin\theta, \quad P^{(s)}_{2,1}(\theta) = s(s+2)\sin\theta\cos\theta. \tag{3.17}$$

Then employing (3.14) and (3.17),

$$\sin\theta\, P^{(s)}_{0,0}(\theta) = (1/s)P^{(s)}_{1,1}(\theta), \tag{3.18a}$$

$$\sin\theta\, P^{(s)}_{1,0}(\theta) = [1/(s+2)]P^{(s)}_{2,1}(\theta). \tag{3.18b}$$

Here, and in all subsequent equations, we adopt the convention $P^{(s)}_{p,r} = 0$ if $p < r$. Then (3.16) includes (3.18) as special cases and is therefore valid for all $n = 0,1,2,\ldots$. The identities (3.13) and (3.16) allow (3.12) to be written in the form

$$\begin{aligned}
&|p|(|p|^2 + M^2)f^{(\pm)}(|p|)[-i(P^{(2)}_{n+1,0}(\theta_2) + P^{(2)}_{n-1,0}(\theta_2))\,\phi^{(+)}_{1/2,m}(\theta_3,\phi) \\
&\quad \pm [1/(n+1)](P^{(2)}_{n+1,1}(\theta_2) - P^{(2)}_{n-1,1}(\theta_2))\,\phi^{(-)}_{1/2,m}(\theta_3,\phi)] + 2m(|p|^2 + M^2)f^{(\mp)}(|p|)P^{(2)}_{n,0}(\theta_2)\phi^{(+)}_{1/2,m}(\theta_3,\phi) \\
&\quad = -\frac{qQ}{(2\pi)^4}(-ip^0 \pm \boldsymbol{\sigma}\cdot\mathbf{p})\int \frac{d\,|q|d\Omega'_{(3)}\,|q|^3}{|p|^2 + |q|^2 - 2|p|\;|q|\cos\Theta_{(4)}}\,2f^{(\pm)}(|q|)P^{(2)}_{n,0}(\theta'_2)\phi^{(+)}_{1/2,m}(\theta'_3,\phi') \\
&\quad\quad -\frac{qQ}{(2\pi)^4}\int \frac{d\,|q|d\Omega'_{(3)}\,|q|^4}{|p|^2 + |q|^2 - 2|p|\;|q|\cos\Theta_{(4)}}f^{(\pm)}(|q|)[-i(P^{(2)}_{n+1,0}(\theta'_2) + P^{(2)}_{n-1,0}(\theta'_2))\phi^{(+)}(\theta'_3,\phi') \\
&\quad\quad \pm [1/(n+1)](P^{(2)}_{n+1,1}(\theta'_2) - P^{(2)}_{n-1,1}(\theta'_2))\phi^{(-)}_{1/2,m}(\theta'_3,\phi')].
\end{aligned} \tag{3.19}$$

The two integrals over $d\Omega'_{(3)}$ can be performed using Hecke's theorem.[9] Adopting the notation of Ref. 4, Appendix A, the above equation takes the form

$$\begin{aligned}
&|p|(|p|^2 + M^2)f^{(\pm)}(|p|)[-i(P^{(2)}_{n+1,0}(\theta_2) + P^{(2)}_{n-1,0}(\theta_2))\phi^{(+)}_{1/2,m}(\theta_3,\phi) \\
&\quad \pm [1/(n+1)](P^{(2)}_{n+1,1}(\theta_2) - P^{(2)}_{n-1,1}(\theta_2))\phi^{(-)}_{1/2,m}(\theta_3,\phi)] + 2m(|p|^2 + M^2)f^{(\mp)}(|p|)P^{(2)}_{n,0}(\theta_2)\phi^{(+)}_{1/2,m}(\theta_3,\phi) \\
&\quad = -[qQ/(2\pi)^4]|p|[-i(P^{(2)}_{n+1,0}(\theta_2) + P^{(2)}_{n-1,0}(\theta_2))\phi^{(+)}_{1/2,m}(\theta_3,\phi) \pm [1/(n+1)](P^{(2)}_{n+1,1}(\theta_2) \\
&\quad\quad - P^{(2)}_{n-1,1}(\theta_2))\phi^{(-)}_{1/2,m}(\theta_3,\phi)]\int d\,|q|\;|q|^3 f^{(\pm)}(|q|)\Lambda_n(|p|,|q|) \\
&\quad\quad -\frac{qQ}{(2\pi)^4}\left[-i\,P^{(2)}_{n+1,0}(\theta_2)\phi^{(+)}(\theta_3,\phi) \pm \frac{1}{n+1}P^{(2)}_{n+1,1}(\theta_2)\phi^{(-)}_{1/2,m}(\theta_3,\phi)\right]\int d\,|q|\;|q|^4 f^{(\pm)}(|q|)\Lambda_{n+1}(|p|,|q|) \\
&\quad\quad -\frac{qQ}{(2\pi)^4}\left[-iP^{(2)}_{n-1,0}(\theta_2)\phi^{(+)}(\theta_3,\phi) \mp \frac{1}{n+1}P^{(2)}_{n-1,1}(\theta_2)\phi^{(-)}_{1/2,m}(\theta_3,\phi)\right]\int d\,|q|\;|q|^4 f^{(\pm)}(|q|)\Lambda_{n-1}(|p|,|q|),
\end{aligned} \tag{3.20}$$

where

$$\Lambda_n(|p|,|q|) = \frac{4\pi}{n+1}\int_{-1}^{1}\frac{dx}{|p|^2 + |q|^2 - 2|p|\;|q|x} \times\sqrt{1-x^2}\,C^1_n(x). \tag{3.21}$$

In obtaining (3.20), after evaluating the first integral in (3.19), identities (3.13) and (3.16) were used. In (3.21), $C^1_n(x)$ is a Gegenbauer polynomial. The integral in (3.21) can be performed with the aid of a table of integrals[12] yielding

$$\Lambda_n(|p|,|q|) = \frac{-i(2\pi)^{3/2}}{n+1}\,\frac{(\xi^2-1)^{1/4}}{|p|\;|q|}\,Q^{1/2}_{n+1/2}(\xi), \tag{3.22}$$

where $\xi = (|p|^2 + |q|^2)/2|p|\;|q|$ and $Q^{1/2}_{n+1/2}$ is an associated Legendre function of the second kind.

The hyperspherical harmonics are linearly independent so the coefficient of $P^{(2)}_{n,0}(\theta_2)\phi^{(+)}_{1/2,m}(\theta_3,\phi)$ in (3.20) must vanish yielding

$$2m(|p|^2 + M^2)f^{(\mp)}(|p|) = 0. \tag{3.23}$$

Since the above equation must be satisfied for all $|p|$, the fermion mass m must equal zero if solutions of the form (3.9) exist. But even if $m = 0$, the angular dependence does not necessarily separate because, from (3.22), $\Lambda_{n-1}(|p|,|q|) \neq \Lambda_{n+1}(|p|,|q|)$. However, recalling the convention $P^{(s)}_{p,r} = 0$ if $p < r$, if the fermion mass $m = 0$ and the index $n = 0$, then the angular dependence of the solutions in (3.20) does separate. If we had included the seagull term, we would have arrived at the same conclusions, but the calculation is more involved. Thus when we solve the Bethe–Salpeter equation in the ladder approximation or when the seagull contribution is also included, we require the constituent fermion mass $m=0$ and seek $l=0$ eigenstates of the form

$$\chi^{(\pm)}(|p|)=2f^{(\pm)}(|p|)P^{(s)}_{0,0}(\theta_2)\phi^{(+)}_{j=1/2,m}(\theta_3,\phi). \qquad (3.24)$$

IV. ZERO ORBITAL ANGULAR MOMENTUM, ZERO FOUR-MOMENTUM HELICITY EIGENSTATES OF THE BETHE–SALPETER EQUATION IN THE LADDER APPROXIMATION

To solve the Bethe–Salpeter equation, we use the method of Fock[6–8] and project four-dimensional momentum space onto the surface of a five-dimensional hypersphere with the transformation

$$|p|=M\tan(\theta_1/2). \qquad (4.1)$$

The factor 2 is included in the above formula because the range of θ_1 on a hypersphere must be $0\leqslant\theta_1\leqslant\pi$ so as θ_1 varies over this range, $|p|$ varies from 0 to ∞ as required. Defining the four-vector q in analogy with (3.6) and (4.1) except that the angles are denoted by primes,

$$d^4q=[M^4/16\cos^8(\theta'_1/2)]\sin^3\theta'_1\sin^2\theta'_2\sin\theta'_3 \times d\theta'_1\,d\theta'_2\,d\theta'_3\,d\phi', \qquad (4.2a)$$

$$\equiv[M^4/16\cos^8(\theta'_1/2)]d\Omega'_{(4)}. \qquad (4.2b)$$

The components of the unit vector $\hat{u}_{(5)}$ in five dimensions are

$$\hat{u}_{(5)}=(\cos\theta_1,\sin\theta_1\cos\theta_2, \sin\theta_1\sin\theta_2\cos\theta_3,\ \sin\theta_1\sin\theta_2\sin\theta_3\cos\phi, \sin\theta_1\sin\theta_2\sin\theta_3\sin\phi). \qquad (4.3)$$

Using a corresponding expression for the unit vector $\hat{v}_{(5)}$ in terms of primed angles, it is straightforward to show that

$$(p-q)\cdot(p-q) =[M^2/\cos^2(\theta_1/2)\cos^2(\theta'_1/2)]\tfrac{1}{2}(1-\cos\Theta). \qquad (4.4)$$

In (4.4), Θ is the angle between the unit vectors $\hat{u}_{(5)}$ and $\hat{v}_{(5)}$. Setting $m=0$ and using (4.2) and (4.4), (3.5) becomes

$$\begin{aligned}(-ip^0\pm\sigma^ip^i)M^2(1+\tan^2(\theta_1/2))\chi^{(\pm)}(p) &=-\frac{qQ}{(2\pi)^4}\left\{(-ip^0\pm\sigma^ip^i)\int d\Omega'_{(4)}\frac{M^4}{16\cos^8(\theta'_1/2)}\right.\\ &\quad\times\frac{2\cos^2(\theta_1/2)\cos^2(\theta'_1/2)}{M^2(1-\cos\Theta)}\chi^{(\pm)}(q)+\int d\Omega'_{(4)}\\ &\quad\times\frac{M^4}{16\cos^8(\theta'_1/2)}\frac{2\cos^2(\theta_1/2)\cos^2(\theta'_1/2)}{M^2(1-\cos\Theta)}\\ &\quad\left.\times(-iq^0\pm\sigma^iq^i)\chi^{(\pm)}(q)\right\}. \end{aligned} \qquad (4.5)$$

We seek $l=0$ solutions of the form (3.24). To simplify the above equation we note that

$$\begin{aligned}(-ip^0\pm\sigma^ip^i)\chi^{(\pm)}(p) &=f^{(\pm)}(|p|)[-i|p|\cos\theta_2\,2P^{(2)}_{0,0}(\theta_2)\phi^{(+)}_{1/2,m}(\theta_3,\phi)\\ &\quad\pm|p|\sin\theta_2\,2P^{(2)}_{0,0}(\theta_2)\phi^{(-)}_{1/2,m}(\theta_3,\phi)]. \end{aligned} \qquad (4.6)$$

With the aid of (3.15) and (3.18a), the above equation becomes

$$\begin{aligned}(-ip^0\pm\sigma^ip^i)\chi^{(\pm)}(p) &=f^{(\pm)}(|p|)|p|[-iP^{(2)}_{1,0}(\theta_2)\phi^{(+)}_{1/2,m}(\theta_3,\phi)\\ &\quad\pm P^{(2)}_{1,1}(\theta_2)\phi^{(-)}_{1/2,m}(\theta_3,\phi)]. \end{aligned} \qquad (4.7)$$

Defining

$$\tilde{\psi}_0(\theta_2,\theta_3,\phi)\equiv P^{(2)}_{0,0}(\theta_2)\phi^{(+)}_{1/2,m}(\theta_3,\phi), \qquad (4.8a)$$

$$\tilde{\psi}^{(\pm)}_1(\theta_2,\theta_3,\phi)\equiv-iP^{(2)}_{1,0}(\theta_2)\phi^{(+)}_{1/2,m}(\theta_3,\phi) \pm P^{(2)}_{1,1}(\theta_2)\phi^{(-)}_{1/2,m}(\theta_3,\phi), \qquad (4.8b)$$

(4.7) becomes

$$(-ip^0\pm\sigma^ip^i)\tilde{\psi}_0(\theta_2,\theta_3,\phi)=(|p|/2)\tilde{\psi}^{(\pm)}_1(\theta_2,\theta_3,\phi). \qquad (4.9)$$

With the above results, (4.5) takes the form

$$\begin{aligned}M^3\frac{\sin(\theta_1/2)}{\cos^3(\theta_1/2)}f^{(\pm)}(|p|)\tilde{\psi}^{(\pm)}_1(\theta_2,\theta_3\phi) &=-\frac{qQ}{(2\pi)^4}\frac{M^2}{8}\cos^2\frac{\theta_1}{2}\left[2(-ip^0\pm\sigma^ip^i)\right.\\ &\quad\times\int\frac{d\Omega'_{(4)}}{1-\cos\Theta}\frac{\cos^2(\theta'_1/2)}{\cos^8(\theta'_1/2)}f^{(\pm)}(|q|)\\ &\quad\times\tilde{\psi}_0(\theta'_2,\theta'_3,\phi')+\int\frac{d\Omega'_{(4)}}{1-\cos\Theta}\frac{\cos^2(\theta'_1/2)}{\cos^8(\theta'_1/2)}M\\ &\quad\left.\times\tan(\theta'_1/2)f^{(\pm)}(|q|)\tilde{\psi}^{(\pm)}_1(\theta'_2,\theta'_3,\phi')\right]. \end{aligned} \qquad (4.10)$$

The function $f^{(\pm)}(|p|)$ is assumed to be of the form

$$\frac{f^{(\pm)}(|p|)}{\cos^8(\theta_1/2)}=\sum_{n=0}^{\infty}(2n+2\nu+3)f^{(\pm)}_nP^{(3)}_{n+\nu,0}(\theta_1), \quad \nu=0,1,2,..., \qquad (4.11)$$

where the $f^{(\pm)}_n$ are constants and the factor of $2n+2\nu+3$ has been included for computational convenience. Substituting (4.11) into (4.10), the Bethe–Salpeter equation in the ladder approximation becomes

$$\begin{aligned}M^3\sin\frac{\theta_1}{2}\cos^5\frac{\theta_1}{2} &\times\sum_{n=0}^{\infty}(2n+2\nu+3)f^{(\pm)}_nP^{(3)}_{n+\nu,0}(\theta_1)\tilde{\psi}^{(\pm)}_1(\theta_2,\theta_3,\phi)\\ &=-\frac{qQ}{(2\pi)^4}\frac{M^2}{8}(I_1+I_2), \end{aligned} \qquad (4.12)$$

where

$$\begin{aligned}I_1=(-ip^0\pm\sigma^ip^i)\cos^2\frac{\theta_1}{2}\int\frac{d\Omega'_{(4)}}{1-\cos\Theta}2\cos^2\frac{\theta'_1}{2} &\times\sum_{n=0}^{\infty}(2n+2\nu+3)f^{(\pm)}_nP^{(3)}_{n+\nu,0}(\theta'_1)\tilde{\psi}_0(\theta'_2,\theta'_3,\phi'), \end{aligned} \qquad (4.13)$$

and

$$I_2 = M\cos^2\frac{\theta_1}{2}\int\frac{d\Omega'_{(4)}}{1-\cos\Theta}\sin\frac{\theta'_1}{2}\cos\frac{\theta'_1}{2}$$

$$\times\sum_{n=0}^{\infty}(2n+2\nu+3)f_n^{(\pm)}P^{(3)}_{n+\nu,0}(\theta'_1)$$

$$\times\tilde{\psi}_1^{(\pm)}(\theta'_2,\theta'_3,\phi'). \tag{4.14}$$

The left-hand side (lhs) of (4.12) and the integrals I_1 and I_2 will now be evaluated one at a time with $\nu = 0$. The $\nu = 1$ solution can be obtained from the $\nu = 0$ solution simply by setting $f_0^{(\pm)} = 0$. The $\nu = 2,3,...$ solutions can be obtained from the $\nu = 0$ solution in a similar fashion.

Using the trigonometric identities $[\sin(\theta/2)][\cos(\theta/2)] = \frac{1}{2}\sin\theta$ and $\cos^2(\theta/2) = (\frac{1}{2})(1+\cos\theta)$ as well as identities in the Appendix of Ref. 4,

lhs of (4.12)

$$= \frac{M^3}{8}\sum_{n=1}^{\infty}\Bigg\{\frac{(n-2)(n-1)}{(2n-1)(2n+1)}f^{(\pm)}_{n-3} + \frac{2(n-1)}{2n+1}f^{(\pm)}_{n-2}$$

$$+\left[1-\frac{(n-1)(n-2)}{(2n-1)(2n+1)}+\frac{(n-1)(n+3)}{(2n+1)(2n+3)}+\frac{n(n+4)}{(2n+3)(2n+5)}\right]f^{(\pm)}_{n-1}+\frac{6(2n+3)}{(2n+1)(2n+5)}f_n^{(\pm)}$$

$$-\left[1+\frac{(n-1)(n+3)}{(2n+1)(2n+3)}+\frac{n(n+4)}{(2n+3)(2n+5)}-\frac{(n+4)(n+5)}{(2n+5)(2n+7)}\right]f^{(\pm)}_{n+1}$$

$$-\frac{2(n+4)}{2n+5}f^{(\pm)}_{n+2}-\frac{(n+4)(n+5)}{(2n+5)(2n+7)}f^{(\pm)}_{n+3}\Bigg\}P^{(3)}_{n,1}(\theta_1)\tilde{\psi}_1^{(\pm)}(\theta_2,\theta_3,\phi). \tag{4.15}$$

In deriving the above equation, the identities in the Appendix of Ref. 4 were only used when $p \geqslant r$ for all $P^{(s)}_{p,r}$ appearing in the identity. If one of the $P^{(s)}_{p,r}$ did not satisfy this condition, then the explicit expressions for the $P^{(s)}_{p,r}$ of interest were used to simplify (4.15).

To evaluate the integral I_1 we set $\nu = 0$ and use $2\cos^2(\theta'_1/2) = (1+\cos\theta'_1)$ yielding

$$I_1 = (-ip^0 \pm \sigma^i p^i)\cos^2\frac{\theta_1}{2}\int\frac{d\Omega'_{(4)}}{1-\cos\Theta}(1+\cos\theta'_1)\sum_{n=0}^{\infty}(2n+3)f_n^{(\pm)}P^{(3)}_{n,0}(\theta'_1)\tilde{\psi}_0(\theta'_2,\theta'_3,\phi'). \tag{4.16}$$

From (3.13)

$$I_1 = (-ip^0 \pm \sigma^i p^i)\cos^2\frac{\theta_1}{2}\int\frac{d\Omega'_{(4)}}{1-\cos\Theta}$$

$$\times\left\{\sum_{n=0}^{\infty}(2n+3)f_n^{(\pm)}P^{(3)}_{n,0}(\theta'_1)+\sum_{n=0}^{\infty}f_n^{(\pm)}[(n+2)P^{(3)}_{n-1,0}(\theta'_1)+(n+1)P^{(3)}_{n+1,0}(\theta'_1)]\right\}\tilde{\psi}_0(\theta'_2,\theta'_3,\phi'). \tag{4.17}$$

Each term under the integral is a hyperspherical harmonic and can therefore be integrated using Hecke's theorem.[9] Using the notation of Ref. 4 [see Eqs. (A9) and (A10)], after integration (4.17) becomes

$$I_1 = (-ip^0 \pm \sigma^i p^i)\cos^2\frac{\theta_1}{2}$$

$$\times\sum_{n=0}^{\infty}f_n^{(\pm)}[(2n+3)\Lambda_n P^{(3)}_{n,0}(\theta_1)+(n+2)\Lambda_{n-1}P^{(3)}_{n-1,0}(\theta_1)+(n+1)\Lambda_{n+1}P^{(3)}_{n+1,0}(\theta_1)]\tilde{\psi}_0(\theta_2,\theta_3,\phi), \tag{4.18}$$

where

$$\Lambda_n = \frac{4\pi^2}{(n+2)(n+1)}\int_{-1}^{1}dx(1+x)C_n^{3/2}(x)dx = \frac{8\pi^2}{(n+2)(n+1)}. \tag{4.19}$$

The integral in (4.19) is easily evaluated employing the technique discussed in the Appendix of Ref. 4. Using (4.9), (4.1), and the identity $[\sin(\theta_1/2)][\cos(\theta_1/2)] = \frac{1}{2}\sin\theta_1$,

$$I_1 = \frac{M}{4}\sin\theta_1\sum_{n=0}^{\infty}f_n^{(\pm)}[(2n+3)\Lambda_n P^{(3)}_{n,0}(\theta_1)+(n+2)\Lambda_{n-1}P^{(3)}_{n-1,0}(\theta_1)+(n+1)\Lambda_{n+1}P^{(3)}_{n+1,0})(\theta_1)]\tilde{\psi}_1^{(\pm)}(\theta_2,\theta_3,\phi). \tag{4.20}$$

Since (3.16) has been established for all $n = 0,1,2,...$,

$$I_1 = \frac{M}{4}\sum_{n=0}^{\infty}f_n^{(\pm)}\Bigg\{\Lambda_n[P^{(3)}_{n+1,1}(\theta_1)-P^{(3)}_{n-1,1}(\theta_1)]+\frac{n+2}{2n+1}\Lambda_{n-1}[P^{(3)}_{n,1}(\theta_1)-P^{(3)}_{n-2,1}(\theta_1)]$$

$$+\frac{n+1}{2n+5}\Lambda_{n+1}[P^{(3)}_{n+2,1}(\theta_1)-P^{(3)}_{n,1}(\theta_2)]\Bigg\}\tilde{\psi}_1^{(\pm)}(\theta_2,\theta_3,\phi), \tag{4.21}$$

$$
= \frac{M}{4} \sum_{n=1}^{\infty} \Big[\frac{n-1}{2n+1} \Lambda_{n-1} f_{n-2}^{(\pm)} + \Lambda_{n-1} f_{n-1}^{(\pm)} + \Big(\frac{n+2}{2n+1} \Lambda_{n-1} - \frac{n+1}{2n+5} \Lambda_{n+1} \Big) f_n^{(\pm)}
$$

$$
- \Lambda_{n+1} f_{n+1}^{(\pm)} - \frac{n+4}{2n+5} \Lambda_{n+1} f_{n+2}^{(\pm)} \Big] P_{n,1}^{(3)}(\theta_1) \tilde{\psi}_1^{(\pm)}(\theta_2,\theta_3,\phi). \qquad (4.22)
$$

The integral I_2 is evaluated in a similar fashion and is given by

$$
I_2 = \frac{M}{4} \sum_{n=1}^{\infty} \Big[\frac{n-1}{2n+1} \Lambda_{n-1} f_{n-2}^{(\pm)} + \Lambda_n f_{n-1}^{(\pm)} - \Big(\frac{n-1}{2n+1} \Lambda_{n-1} - \frac{n+4}{2n+5} \Lambda_{n+1} \Big) f_n^{(\pm)}
$$

$$
- \Lambda_n f_{n+1}^{(\pm)} - \frac{n+4}{2n+5} \Lambda_{n+1} f_{n+2}^{(\pm)} \Big] P_{n,1}^{(3)}(\theta_1) \tilde{\psi}_1^{(\pm)}(\theta_2,\theta_3,\phi). \qquad (4.23)
$$

With the aid of the explicit expression (4.19) for Λ_n, the sum of the integrals I_1 and I_2 is

$$
I_1 + I_2 = 2\pi^2 M \sum_{n=1}^{\infty} \Big\{ \frac{2(n-1)}{(2n+1)n(n+1)} f_{n-2}^{(\pm)} + \frac{2}{n(n+2)} f_{n-1}^{(\pm)} + 3 \Big[\frac{1}{n(n+1)(2n+1)} + \frac{1}{(n+2)(n+3)(2n+5)} \Big]
$$

$$
\times f_n^{(\pm)} - \frac{2}{(n+1)(n+3)} f_{n+1}^{(\pm)} - \frac{2(n+4)}{(n+2)(n+3)(2n+5)} f_{n+2}^{(\pm)} \Big\} P_{n,1}^{(3)}(\theta_1) \tilde{\psi}_1^{(\pm)}(\theta_2,\theta_3,\phi). \qquad (4.24)
$$

Substituting (4.15) and (4.24) into (4.12) and canceling common factors, the $f_n^{(\pm)}$ must satisfy the following set of equations:

$$
\frac{(n-2)(n-1)}{(2n-1)(2n+1)} f_{n-3}^{(\pm)} + \frac{2(n-1)}{2n+1} f_{n-2}^{(\pm)}
$$

$$
+ \Big[1 - \frac{(n-2)(n-1)}{(2n-1)(2n+1)} + \frac{(n-1)(n+3)}{(2n+1)(2n+3)} \Big] + \frac{n(n+4)}{(2n+3)(2n+5)} \Big] f_{n-1}^{(\pm)} + \frac{6(2n+3)}{(2n+1)(2n+5)} f_n^{(\pm)}
$$

$$
- \Big[1 + \frac{(n-1)(n+3)}{(2n+1)(2n+3)} + \frac{n(n+4)}{(2n+3)(2n+5)} - \frac{(n+4)(n+5)}{(2n+5)(2n+7)} \Big] f_{n+1}^{(\pm)}
$$

$$
- \frac{2(n+4)}{2n+5} f_{n+2}^{(\pm)} - \frac{(n+4)(n+5)}{(2n+5)(2n+7)} f_{n+3}^{(\pm)}
$$

$$
= - \frac{qQ}{8\pi^2} \Big\{ \frac{2(n-1)}{(2n+1)n(n+1)} f_{n-2}^{(\pm)} + \frac{2}{n(n+2)} f_{n-1}^{(\pm)} + 3 \Big[\frac{1}{n(n+1)(2n+1)} + \frac{1}{(n+2)(n+3)(2n+5)} \Big] f_n^{(\pm)}
$$

$$
- \frac{2}{(n+1)(n+3)} f_{n+1}^{(\pm)} - \frac{2(n+4)}{(n+2)(n+3)(2n+5)} f_{n+2}^{(\pm)} \Big\}, \quad n, = 1,2,\ldots,\infty. \qquad (4.25)
$$

Note that the equations for $f_n^{(+)}$ are identical to those for $f_n^{(-)}$ so we drop the superscript and write $f_n^{(\pm)} \equiv f_n$. The $n = 1$ equation allows f_4 to be expressed in terms of f_0, f_1, f_2, and f_3. Continuing in this manner, all f_n for $n \geqslant 4$ can be expressed in terms f_0, f_1, f_2, and f_3. We thus obtain four linearly independent solutions for any value of the coupling constant.

Here solutions are obtained for any value of the coupling constant while in Ref. 4 solutions were obtained only for specific discrete values. This significant difference in the solutions arises because of the difference in the way the equations separate. In Ref. 4 there are two possibilities for the angular dependence, $\psi_1^{(\pm)}(\theta_2,\theta_3,\phi)$ or $\psi_2^{(\pm)}(\theta_2,\theta_3,\phi)$. The coefficients of both $\psi_1^{(\pm)}(\theta_2,\theta_3,\phi)$ and $\psi_2^{(\pm)}(\theta_2,\theta_3,\phi)$ must vanish yielding two sets of equations. Consistency between the two sets of equations then leads to the requirement that the coupling constant equals one of a specific discrete set of eigenvalues. In contrast, for the $l = 0$ solutions discussed here, the possible angular dependence is given by the single function $\psi_0(\theta_2,\theta_3,\phi)$ defined in (4.8a). Only one set of equations is obtained so there is no consistency equation and no resulting restriction on the possible values of the coupling constant.

V. ZERO ORBITAL ANGULAR MOMENTUM, ZERO FOUR-MOMENTUM HELICITY EIGENSTATES INCLUDING CONTRIBUTIONS FROM BOTH SINGLE PHOTON EXCHANGE AND THE SEAGULL DIAGRAM

Using the same procedures that led to (4.12), when the contribution of the seagull diagram is included, the Bethe–Salpeter equation takes the form

$$
M^3 \sin \frac{\theta_1}{2} \cos^5 \frac{\theta_1}{2} \sum_{n=0}^{\infty} (2n + 2\nu + 3)
$$

$$
\times f_n^{(\pm)} P_{n+\nu,0}^{(3)}(\theta_1) \tilde{\psi}_1^{(\pm)}(\theta_2,\theta_3,\phi)
$$

$$
= - \frac{qQ}{(2\pi)^4} \frac{M^2}{8} (I_1 + I_2) + \frac{q^2 Q^2}{(2\pi)^8} \frac{M^2}{16} I_3. \qquad (5.1)
$$

The integrals I_1 and I_2 are given by (4.13) and (4.14), respectively, and I_3 is the eightfold integral

$$
I_3 = \cos^2 \frac{\theta_1}{2} \int \frac{d\Omega'_{(4)}}{1 - \cos \Theta}
$$

$$
\times \frac{1}{\cos^2(\theta'_1/2) \sin^2(\theta'_1/2)} (-iq^0 \pm \sigma^i q^i)
$$

$$\times\int\frac{d\Omega''_{(4)}}{1-\cos\Theta'}\,2\cos^2\frac{\theta''_1}{2}\sum_{n=0}^{\infty}(2n+2\nu+3)$$
$$\times f_n^{(\pm)}P^{(3)}_{n+\nu,0}(\theta''_1)\tilde{\psi}_0(\theta''_2,\theta''_3,\phi''). \qquad (5.2)$$

The integration variable k appearing in (2.9) has been expressed in the above equation in terms of double-primed angles by introducing polar coordinates in analogy with (3.6) and (4.1). The angle Θ' is the angle between the unit vectors $\hat{v}_{(5)}$ and $\hat{w}_{(5)}$, where $\hat{v}_{(5)}$ and $\hat{w}_{(5)}$ are given by (4.3) except that the unprimed angles are replaced by primed or double-primed angles, respectively.

Without loss of generality, we set $\nu=0$ as discussed in the previous section. Then using $\cos^2(\theta''_1/2)=\frac{1}{2}(1+\cos\theta''_1)$,

$$I_3=\cos^2\frac{\theta_1}{2}\int\frac{d\Omega'_{(4)}}{1-\cos\Theta}\,\frac{1}{\cos^2(\theta'_1/2)\sin^2(\theta'_1/2)}$$
$$\times(-iq^0\pm\sigma^i q^i)\int\frac{d\Omega''_{(4)}}{1-\cos\Theta'}(1+\cos\theta''_1)$$
$$\times\sum_{n=0}^{\infty}(2n+3)f_n^{(\pm)}P^{(3)}_{n,0}(\theta''_1)\tilde{\psi}_0(\theta''_2,\theta''_3,\phi''). \qquad (5.3)$$

With the aid of the identity (3.13) which has been established for all integer values of $n\geqslant 0$,

$$I_3=\cos^2\frac{\theta_1}{2}\int\frac{d\Omega'_{(4)}}{1-\cos\Theta}\,\frac{1}{\cos^2(\theta'_1/2)\sin^2(\theta'_1/2)}$$
$$\times(-iq^0\pm\sigma^i q^i)\int\frac{d\Omega''_{(4)}}{1-\cos\Theta'}\sum_{n=0}^{\infty}f_n^{(\pm)}$$
$$\times[(2n+3)P^{(3)}_{n,0}(\theta''_1)+(n+1)P^{(3)}_{n+1,0}(\theta''_1)$$
$$+(n+2)P^{(3)}_{n-1,0}(\theta''_1)]\tilde{\psi}_0(\theta''_2,\theta''_3,\phi''). \qquad (5.4)$$

The integral over $d\Omega''_{(4)}$ can now be evaluated using Hecke's theorem,[4,9]

$$I_3=\cos^2\frac{\theta_1}{2}\int\frac{d\Omega'_{(4)}}{1-\cos\Theta}\,\frac{1}{\cos^2(\theta'_1/2)\sin^2(\theta'_1/2)}$$
$$\times(-iq^0\pm\sigma^i q^i)\sum_{n=0}^{\infty}f_n^{(\pm)}$$
$$\times[(2n+3)\Lambda_nP^{(3)}_{n,0}(\theta'_1)+(n+1)\Lambda_{n+1}P^{(3)}_{n+1,0}(\theta'_1)$$
$$+(n+2)\Lambda_{n-1}P^{(3)}_{n-1,0}(\theta'_1)]\tilde{\psi}_0(\theta'_2,\theta'_3,\phi'), \qquad (5.5)$$

where Λ_n is given in (4.19). The product $(-iq^0\pm\sigma^iq^i)\tilde{\psi}_0(\theta'_2,\theta'_3,\phi')$ follows immediately from (4.9) yielding

$$I_3=\frac{M}{2}\cos^2\frac{\theta_1}{2}\int\frac{d\Omega'_{(4)}}{1-\cos\Theta}\,\frac{1}{\cos^3(\theta'_1/2)\sin(\theta'_1/2)}$$
$$\times\sum_{n=0}^{\infty}f_n^{(\pm)}[(2n+3)\Lambda_nP^{(3)}_{n,0}(\theta'_1)$$
$$+(n+1)\Lambda_{n+1}P^{(3)}_{n+1,0}(\theta'_1)$$
$$+(n+2)\Lambda_{n-1}P^{(3)}_{n-1,0}(\theta'_1)]\tilde{\psi}_1^{(\pm)}(\theta'_2,\theta'_3,\phi'). \qquad (5.6)$$

Expressing $\cos(\theta/2)$ and $\sin(\theta/2)$ in terms of $\cos\theta$ and $\sin\theta$,

$$I_3=M(1+\cos\theta_1)\int\frac{d\Omega'_{(4)}}{1-\cos\Theta}\,\frac{1}{(1+\cos\theta'_1)\sin\theta'_1}$$
$$\times\sum_{n=0}^{\infty}f_n^{(\pm)}[(2n+3)\Lambda_nP^{(3)}_{n,0}(\theta'_1)$$
$$+(n+1)\Lambda_{n+1}P^{(3)}_{n+1,0}(\theta'_1)$$
$$+(n+2)\Lambda_{n-1}P^{(3)}_{n-1,0}(\theta'_1)]\tilde{\psi}_1^{(\pm)}(\theta'_2,\theta'_3,\phi'). \qquad (5.7)$$

The above integral cannot be evaluated immediately because of the factor $[(1+\cos\theta'_1)\sin\theta'_1]^{-1}$. However, if we choose the constants $f_n^{(\pm)}$ such that the sum on the right-hand side of the above equation is given by

$$\sum_{n=0}^{\infty}f_n^{(\pm)}[(2n+3)\Lambda_nP^{(3)}_{n,0}(\theta'_1)$$
$$+(n+1)\Lambda_{n+1}P^{(3)}_{n+1,0}(\theta'_1)$$
$$+(n+2)\Lambda_{n-1}P^{(3)}_{n-1,0}(\theta'_1)]$$
$$=(1+\cos\theta'_1)\sin\theta'_1\sum_{n=1}^{\infty}g_n^{(\pm)}P^{(3)}_{n,1}(\theta'_1), \qquad (5.8)$$

where the $g_n^{(\pm)}$ are constants, the integral in (5.7) can be evaluated using Hecke's theorem.[4,9] Our strategy, then, is to evaluate I_3 in terms of $g_n^{(\pm)}$. When the expression for I_3 and results from Sec. IV are substituted into the Bethe–Salpeter equation (5.1), the equation becomes a recursion relation for the $f_n^{(\pm)}$ and $g_n^{(\pm)}$ which are not all independent. We then solve (5.8) and express the $f_n^{(\pm)}$ in terms of the $g_n^{(\pm)}$.

To evaluate (5.7) in terms of the $g_i^{(\pm)}$, we substitute (5.8) into (5.7) and integrate using Hecke's theorem,[4,9]

$$I_3=M(1+\cos\theta_1)\int\frac{d\Omega'_{(4)}}{1-\cos\Theta}$$
$$\times\sum_{n=1}^{\infty}g_n^{(\pm)}P^{(3)}_{n,1}(\theta'_1)\tilde{\psi}_1^{(\pm)}(\theta'_2\theta'_3,\phi'),$$
$$=M(1+\cos\theta_1)\sum_{n=1}^{\infty}g_n^{(\pm)}\Lambda_nP^{(3)}_{n,1}(\theta_1)$$
$$\times\tilde{\psi}_1^{(\pm)}(\theta_2,\theta_3,\phi). \qquad (5.9)$$

Making use of Ref. 4 [Eq. (A23)] and the fact that $\cos\theta\times P^{(3)}_{1,1}=\frac{1}{5}P^{(3)}_{2,1}$,

$$I_3=M\Bigg\{\sum_{n=1}^{\infty}g_n^{(\pm)}\Lambda_nP^{(3)}_{n,1}(\theta_1)$$
$$+\sum_{n=1}^{\infty}g_n^{(\pm)}\Lambda_n\frac{1}{2n+3}[(n+3)P^{(3)}_{n-1,1}$$
$$+nP^{(3)}_{n+1,1}]\Bigg\}\tilde{\psi}_1^{(\pm)}(\theta_2,\theta_3,\phi). \qquad (5.10)$$

Changing the summation indices and using the explicit expression (4.19) for Λ_n, we obtain the desired result for I_3,

$$I_3 = 8\pi^2 M \sum_{n=1}^{\infty} \left[\frac{n-1}{n(n+1)(2n+1)} g_{n-1}^{(\pm)} + \frac{1}{(n+1)(n+2)} g_n^{(\pm)} + \frac{n+4}{(n+2)(n+3)(2n+5)} g_{n+1}^{(\pm)}\right] \times P_{n,1}^{(3)}(\theta_1)\tilde{\psi}_1^{(\pm)}(\theta_2,\theta_3,\phi). \qquad (5.11)$$

Substituting (4.15), (4.24), and (5.11) into (5.1), the Bethe–Salpeter equation yields the following recursion relation:

$$\begin{aligned}
&\frac{(n-2)(n-1)}{(2n-1)(2n+1)} f_{n-3}^{(\pm)} + \frac{2(n-1)}{2n+1} f_{n-2}^{(\pm)} \\
&\quad + \left[1 - \frac{(n-2)(n-1)}{(2n-1)(2n+1)} + \frac{(n-1)(n+3)}{(2n+1)(2n+3)} + \frac{n(n+4)}{(2n+3)(2n+5)}\right] f_{n-1}^{(\pm)} \\
&\quad + \frac{6(2n+3)}{(2n+1)(2n+5)} f_n^{(\pm)} - \left[1 + \frac{(n-1)(n+3)}{(2n+1)(2n+3)} + \frac{n(n+4)}{(2n+3)(2n+5)} - \frac{(n+4)(n+5)}{(2n+5)(2n+7)}\right] f_{n+1}^{(\pm)} \\
&\quad - \frac{2(n+4)}{2n+5} f_{n+2}^{(\pm)} - \frac{(n+4)(n+5)}{(2n+5)(2n+7)} f_{n+3}^{(\pm)} \\
&= -\frac{qQ}{8\pi^2}\left\{\frac{2(n-1)}{(2n+1)n(n+1)} f_{n-2}^{(\pm)} + \frac{2}{n(n+2)} f_{n-1}^{(\pm)} + 3\left[\frac{1}{n(n+1)(2n+1)} + \frac{1}{(n+2)(n+3)(2n+5)}\right] f_n^{(\pm)}\right. \\
&\quad \left. - \frac{2}{(n+1)(n+3)} f_{n+1}^{(\pm)} - \frac{2(n+4)}{(n+2)(n+3)(2n+5)} f_{n+2}^{(\pm)}\right\} + \frac{q^2Q^2}{(2\pi)^6} \\
&\quad \times \left[\frac{n-1}{n(n+1)(2n+1)} g_{n-1}^{(\pm)} + \frac{1}{(n+1)(n+2)} g_n^{(\pm)} + \frac{n+4}{(n+2)(n+3)(2n+5)} g_{n+1}^{(\pm)}\right]; \quad n = 1,2,3,\ldots . \qquad (5.12)
\end{aligned}$$

The $f_n^{(\pm)}$ and $g_n^{(\pm)}$ are not independent so (5.8) must now be solved to express $f_n^{(\pm)}$ in terms of the $g_i^{(\pm)}$. Using Eq. (A25) from Ref. 4 and then (3.13), (5.8) becomes

$$\begin{aligned}
&\sum_{n=0}^{\infty} f_n^{(\pm)}\left[(2n+3)\Lambda_n P_{n,0}^{(3)}(\theta_1') + (n+1)\Lambda_{n+1} P_{n+1,0}^{(3)}(\theta_1') + (n+2)\Lambda_{n-1} P_{n-1,0}^{(3)}(\theta_1')\right] \\
&= (1+\cos\theta_1') \sum_{n=1}^{\infty} \frac{g_n^{(\pm)}}{2n+3}\left[(n+3)(n+2)P_{n-1,0}^{(3)}(\theta_1') - n(n+1)P_{n+1,0}^{(3)}(\theta_1')\right] \\
&= \sum_{n=1}^{\infty} \frac{g_n^{(\pm)}}{(2n+3)}\left[(n+3)(n+2)P_{n-1,0}^{(3)}(\theta_1') - n(n+1)P_{n+1,0}^{(3)}(\theta_1')\right] \\
&\quad + \sum_{n=1}^{\infty} \frac{g_n^{(\pm)}(n+3)(n+2)}{(2n+3)(2n+1)}\left[nP_{n,0}^{(3)}(\theta_1') + (n+1)P_{n-2,0}^{(3)}(\theta_1')\right] \\
&\quad - \sum_{n=1}^{\infty} \frac{g_n^{(\pm)}n(n+1)}{(2n+3)(2n+5)}\left[(n+2)P_{n+2,0}^{(3)}(\theta_1') + (n+3)P_{n,0}^{(3)}(\theta_1')\right]. \qquad (5.13)
\end{aligned}$$

Changing the summation indices in various ways and making use of the explicit expression (4.19) for Λ_n, (5.13) can be rewritten in the convenient form

$$\begin{aligned}
&\sum_{n=0}^{\infty}\left[nf_{n-1}^{(\pm)} + (2n+3)f_n^{(\pm)} + (n+3)f_{n+1}^{(\pm)}\right] \frac{8\pi^2}{(n+2)(n+1)} P_{n,0}^{(3)}(\theta_1') \\
&= \sum_{n=0}^{\infty}\left[-\frac{n(n-1)(n-2)}{(2n-1)(2n+1)} g_{n-2}^{(\pm)} - \frac{n(n-1)}{2n+1} g_{n-1}^{(\pm)} + \frac{3n(n+3)}{(2n+1)(2n+5)} g_n^{(\pm)}\right. \\
&\quad \left. + \frac{(n+4)(n+3)}{2n+5} g_{n+1}^{(\pm)} + \frac{(n+5)(n+4)(n+3)}{(2n+7)(2n+5)} g_{n+2}^{(\pm)}\right] P_{n,0}^{(3)}(\theta_1'). \qquad (5.14)
\end{aligned}$$

Since the coefficients of $P_{n,0}^{(3)}(\theta_1')$ on the left-hand and right-hand sides of (5.14) must be equal, the $f_n^{(\pm)}$ and $g_i^{(\pm)}$ must satisfy

$$\begin{aligned}
&\left[nf_{n-1}^{(\pm)} + (2n+3)f_n^{(\pm)} + (n+3)f_{n+1}^{(\pm)}\right] \frac{8\pi^2}{(n+2)(n+1)} \\
&= -\frac{n(n-1)(n-2)}{(2n-1)(2n+1)} g_{n-2}^{(\pm)} - \frac{n(n-1)}{2n+1} g_{n-1}^{(\pm)} + \frac{3n(n+3)}{(2n+1)(2n+5)} g_n^{(\pm)} \\
&\quad + \frac{(n+4)(n+3)}{2n+5} g_{n+1}^{(\pm)} + \frac{(n+5)(n+4)(n+3)}{(2n+7)(2n+5)} g_{n+2}^{(\pm)}; \quad n = 0,1,2,\ldots . \qquad (5.15)
\end{aligned}$$

As can be readily checked, the above equation is satisfied if the $f_n^{(\pm)}$ are given by

$$f_n^{(\pm)} = (-1)^n f_0^{(\pm)} + \frac{n(n+3)}{8\pi^2(2n+3)}\left[-\frac{(n+2)(n-1)}{2n+1}g_{n-1}^{(\pm)} + 2g_n^{(\pm)} + \frac{(n+4)(n+1)}{2n+5}g_{n+1}^{(\pm)}\right]. \tag{5.16}$$

From the above equation we note that if (5.8) is to be satisfied, $f_0^{(\pm)}$ is arbitrary and all $f_n^{(\pm)}$ for $n>0$ are expressed in terms of $f_0^{(\pm)}$ and $g_i^{(\pm)}$, $i\geqslant 1$.

Substituting (5.16) into (5.12), the Bethe–Salpeter equation yields a recursion relation among the $g_i^{(\pm)}$ and $f_0^{(\pm)}$. Just as was the case in the ladder approximation, here the equations for $g_i^{(+)}, f_0^{(+)}$, and $g_i^{(-)}, f_0^{(-)}$ are identical so we drop the superscript. Taking $n=1$ in (5.12) and making use of (5.16), g_5 is determined in terms of f_0, g_1, g_2, g_3, and g_4. Similarly, taking $n=2$, g_6 can be expressed in terms of the same five constants and so forth.

From (3.24), (4.11), and (5.16), the $l=0$ helicity eigenstates of the Bethe–Salpeter equation (5.1) are

$$\begin{aligned}\chi^{(\pm)}(p) &= 2\cos^8\frac{\theta_1}{2}\sum_{n=0}^{\infty}(2n+3)f_n P_{n,0}^{(3)}(\theta_1)P_{0,0}^{(2)}(\theta_2)\phi_{j=1/2,m}^{(+)}(\theta_3,\phi)\\ &= 2\cos^8\frac{\theta_1}{2}\sum_{n=0}^{\infty}\Bigg\{(-1)^n(2n+3)f_0 + \frac{n(n+3)}{8\pi^2}\Bigg[-\frac{(n+2)(n-1)}{2n+1}g_{n-1} + 2g_n\\ &\quad + \frac{(n+4)(n+1)}{2n+5}g_{n+1}\Bigg]\Bigg\}P_{n,0}^{(3)}(\theta_1)P_{0,0}^{(2)}(\theta_2)\phi_{j=1/2,m}^{(+)}(\theta_3,\phi),\end{aligned} \tag{5.17}$$

where f_0 and the g_i satisfy (5.12) and (5.16). Substituting the explicit expressions for g_5, g_6, etc. as determined from (5.12) and (5.16) into (5.17), $\chi^{(\pm)}(p)$ equals a function times f_0 plus a function times g_1 plus $\cdots$ plus a function times g_4. It is straightforward to show that the five functions that, respectively, multiply f_0, g_1, g_2, g_3, and g_4 are linearly independent so there are five linearly independent solutions. Thus when the contribution from the seagull diagram is included in the Bethe–Salpeter equation, the number of zero four-momentum, $l=0$ helicity eigenstates increases from four to five, verifying the importance of the seagull contribution for strongly bound systems. As in the ladder approximation, these solutions exist for all values of the coupling constant.

To determine if this model is actually of physical interest, the form of the Bethe–Salpeter equation that includes the seagull contribution must be solved, both for the finite-energy and lightlike cases. These calculations are currently in progress.

ACKNOWLEDGMENTS

It is a pleasure to thank Professor L. Alayne Parson and Professor Cary Rader for helpful discussions.

[1] G. B. Mainland and D. M. Scott, Nuovo Cimento A **74**, 198 (1983).
[2] J. M. Blatt and V. F. Weiskoff, *Theoretical Nuclear Physics* (Wiley, New York, 1952).
[3] G. B. Mainland and D. M. Scott, Nuovo Cimento A **82**, 357 (1984).
[4] G. B. Mainland, J. Math Phys. **27**, 1344 (1986).
[5] G. C. Wick, Phys. Rev. **96**, 1124 (1954).
[6] V. Fock, Z. Phys. **98**, 145 (1935).
[7] M. Lévy, Proc. R. Soc. London Ser. A **204**, 145 (1950).
[8] R. E. Cutkosky, Phys. Rev. **96**, 1135 (1954).
[9] E. Hecke, Math. Ann. **78**, 398 (1918).
[10] E. E. Salpeter and H. A. Bethe, Phys. Rev. **84**, 1232 (1951).
[11] Our notation is that of J. D. Bjorken and S. D. Drell, *Relativistic Quantum Fields* (McGraw-Hill, New York, 1965). We set $\hbar = c = 1$.
[12] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products* (Academic, New York, 1965). See formula 7.312-1.

Vectorial continued fractions and an algebraic construction of effective Hamiltonians

M. Znojil
Institute of Nuclear Physics, Czechoslovak Academy of Sciences, 250 68 Řež, Czechoslovakia

(Received 16 December 1986; accepted for publication 5 August 1987)

In the haromonic oscillator basis, the effective Hamiltonians are constructed for the polynomial interactions $V(r) = \Sigma_{i=1}^{t} g_i r^{2i}$ and $t \leqslant 4$. Their form (factorized in terms of the vectorial continued fractions) is simpler in a semirelativistic ($c < \infty$) reformulation. Its degeneracy in the $c \to \infty$ limit (confluence of the physical and unphysical solutions) is removed by solving a coupled set of the nonlinear algebraic equations in REDUCE.

I. INTRODUCTION

The general anharmonic oscillator

$$\left(-\frac{d}{dr^2} + \frac{l(l+1)}{r^2} + g_1 r^2 + g_2 r^4 + \cdots + g_t r^{2t} \right) \psi(r) = E\psi(r), \quad g_t > 0, \quad l = (-1), 0, 1, \ldots, \tag{1.1}$$

is a phenomenological as well as methodical laboratory in quantum mechanics.[1] Its use ranges from a purely nonrelativistic description of confinement[2] up to an analysis of the perturbative and nonperturbative aspects of the quantum field theory.[3] In the former case, a maximal numerical efficiency is usually required. The present study of Eq. (1.1) is intended to complement the latter [e.g., Wentzel–Kramers–Brillouin (WKB) or perturbative [4]] type of applications, where an emphasis is laid upon a global and analytic insight into the solutions.

We shall start from the recurrent (matrix continued fractional, MCF,[5] and vectorially continued fractional, VCF[6]) solutions of equations of the type (1.1). In brief, recalling accelerations of their convergence as achieved by a systematic subtraction of the so-called fixed-point (FP) approximants,[7] we shall deliver the explicit nonperturbative approximate forms of the corresponding solvable (finite-dimensional) effective Hamiltonians[8] for $t = 2$, 3, and 4.

First, we describe the recurrent factorization method in its entirely universal form in Sec. II. An essence of both the MCF and (generalized) VCF approaches is shown to lie in an appropriate decomposition of the resolvent. The usual band-matrix assumption (with $H_{mn} = 0$ for all m and n such that $|m-n| > t$) is then found redundant: An arbitrary $t \leqslant \infty$ matrix H is shown to admit the VCF factorization.

The universal $t \leqslant \infty$ VCF method fully preserves a conceptual simplicity of the $t = 1$ special case[9] reviewed briefly in Appendix A. In particular, the close interrelation between the $t = 1$ analytic continued fractional convergence and an asymptotic smoothness of H (Appendix B) is extended to $t > 1$. In Sec. III, this leads to an explicit algebraic nonperturbative definition of the effective Hamiltonians H^{eff}.

In Sec. IV, we apply the general method to the particular bound state problem (1.1). In the first step (Sec. IV A), we recall the results of Ref. 10 and regularize the anharmonicity in Eq. (1.1) by its Fourier symmetrization. Physically, this corresponds to an introduction of some particular relativistic corrections. Methodically this enables us to run efficiently through the basic technicalities. Then, an asymptotic degeneracy of the nonrelativistic limit is found to admit a straightforward removal: We reinterpret the recently developed asymptotic formulas for wave functions[11] as an algebraic ansatz, and demonstrate its suitability by means of the computer symbolic manipulations in REDUCE (Sec. IV B).

Section V is a summary.

II. GENERAL HAMILTONIANS AND THE VECTORIAL CONTINUED FRACTIONS

An introduction of a model-space projector $P = \Sigma_{n=0}^{M} |n\rangle\langle n|$ and a trivial partitioning of the Schrödinger equation

$$\begin{aligned} &P(H-E)(P+Q)|\psi\rangle = 0, \\ &Q(H-E)(Q+P)|\psi\rangle = 0, \quad Q = 1 - P, \end{aligned} \tag{2.1}$$

enable one to eliminate the out-of-the-model-space components of the wave functions,

$$Q\,|\psi\rangle = (E - QHQ)^{-1} QHP\,|\varphi\rangle, \quad |\varphi\rangle = P\,|\psi\rangle. \tag{2.2}$$

The rest of Eq. (2.1) acquires the form of an effective finite-dimensional equation

$$H^{\text{eff}}|\varphi\rangle = E\,|\varphi\rangle, \tag{2.3}$$

where[8]

$$H^{\text{eff}} = PHP + PHQ(E - QHQ)^{-1}QHP. \tag{2.4}$$

Without any loss of generality, we may now consider a factorization

$$H - E = UL, \tag{2.5}$$

where U is an upper triangular matrix or any regular matrix with the property $QU = QUQ$. Similarly, the factor matrix L must be such that $LQ = QLQ$. Then, we may rewrite Eq. (2.4) in an explicitly factorized form

$$\begin{aligned} H^{\text{eff}} - \text{E} &= PULP - PULQ(1/QULQ)QULP \\ &= PULP - PUQLP = PUPLP \end{aligned} \tag{2.6}$$

born by Eq. (2.5). This is our starting point.

For the simple tridiagonal Hamiltonian (A3) (cf. Appendix A), the factors PHQ and QHP contain just one nonzero matrix element. Within any model space, the difference between H and H^{eff} concerns also a single element [cf. (A4) and (A5)],

0022-2488/88/010139-09$02.50

$$H^{\text{eff}} = \begin{pmatrix} A_0 & C_1 & & \\ & \ddots & & \\ & B_{M-2} & A_{M-1} & C_M \\ & & B_{M-1} & G_M \end{pmatrix}, \quad G_M = E + \frac{1}{F_M}. \tag{2.7}$$

Thus we may define H^{eff} by means of the analytic continued fractions.[9] A similar prescription is to be derived now for an entirely general matrix H.

A motivation for such an effort is the following. In the realistic (e.g., many-body) systems with complicated Hamiltonians H and simple trial Lanczos state $|0\rangle$, a tridiagonalization of H achieved by means of the Lanczos prescription[12] (A1) will lead to the complicated basis states $|1\rangle, |2\rangle, \ldots$. Vice versa, the product $H|0\rangle$ will contain a number of the basis states $|1\rangle, |2\rangle, \ldots$ whenever we require their reasonable simplicity,[13]

$$H|0\rangle = |0\rangle A_0 + |1\rangle B_0^{(1)} + |2\rangle B_0^{(2)} + \cdots. \tag{2.8a}$$

Similarly, in a repetition of this procedure,

$$H|1\rangle = |0\rangle C_1^{(1)} + |1\rangle A_1 + |2\rangle B_1^{(1)} + |3\rangle B_1^{(2)} + \cdots \tag{2.8b}$$

a number of the new states will be infinite in principle. In this way, a tridiagonality of H will be lost and we may return, at least partially, to a free choice of the suitable basis states. At most, we may expect that the coefficients $B_k^{(j)}$ will be small for $j \gg 1$.

Formally, we may introduce the vectorial notation

$$(B_k^{(1)}, B_k^{(2)}, \ldots) = \mathbf{B}_k^T, \quad \mathbf{B}_k = \begin{pmatrix} B_k^{(1)} \\ B_k^{(2)} \\ \vdots \end{pmatrix}, \quad |\mathbf{k}\rangle = \begin{pmatrix} |k\rangle \\ |k+1\rangle \\ \vdots \end{pmatrix},$$

generalize Eq. (A1),

$$H|k\rangle = |0\rangle C_1^{(k)} + |1\rangle C_2^{(k-1)} + \cdots + |k-1\rangle C_k^{(1)} + |k\rangle A_k + |\mathbf{k+1}\rangle \cdot \mathbf{B}_k, \quad k = 0,1,\ldots, \tag{2.9}$$

and reinterpret our operator or general matrix H as tridiagonal in a purely formal "vectorial" partitioning

$$H = \begin{pmatrix} A_0 & \mathbf{C}_1^T & & \\ \mathbf{B}_0 & A_1 & \mathbf{C}_2^T & \\ & \mathbf{B}_1 & A_2 & \mathbf{C}_3^T \\ & & & \ddots \end{pmatrix}. \tag{2.10}$$

When we introduce also the tilded diagonal matrices

$$\widetilde{F}_k = \begin{pmatrix} F_k & & \\ & F_{k+1} & \\ & & \ddots \end{pmatrix},$$

auxiliary vectors $\mathbf{U}_k^T = (U_k^{(1)}, U_k^{(2)}, \ldots)$, $\mathbf{L}_{k-1}^T = (L_{k-1}^{(1)}, L_{k-1}^{(2)}, \ldots)$, $k = 1,2,\ldots$, and vectors with superscripts [omitted components, $C_k^{T[m]} = (C_k^{(m+1)}, C_k^{(m+2)}, \ldots)$, $k,m \geqslant 1$], we may write (2.5),

$$H - E = \begin{pmatrix} A_0 - E & \mathbf{C}_1^T & \\ \mathbf{B}_0 & A_1 - E & \mathbf{C}_2^T \\ & & \ddots \end{pmatrix}$$
$$= \begin{pmatrix} 1 & \mathbf{U}_1^T \cdot \widetilde{F}_1 & \\ & & \mathbf{U}_2^T \cdot \widetilde{F}_2 \\ & & \ddots \end{pmatrix} \cdot \widetilde{F}_0^{-1} \cdot \begin{pmatrix} 1 & & \\ \widetilde{F}_1 \cdot \mathbf{L}_0 & 1 & \\ & \widetilde{F}_2 \cdot \mathbf{L}_1 & 1 \\ & & \ddots \end{pmatrix}. \tag{2.11}$$

This prescription becomes an algebraic identity whenever we satisfy the relations

$$C_{k+1}^{(l)} = U_{k+1}^{(l)} + \sum_{j=1}^{\infty} U_{k+1}^{(l+j)} F_{k+l+j} L_{k+l}^{(j)},$$
$$B_k^{(l)} = L_k^{(l)} + \sum_{j=1}^{\infty} U_{k+l+1}^{(j)} F_{k+l+j} L_k^{(l+j)}, \quad l = 1,2,\ldots, \tag{2.12}$$

and the $l = 0$ requirements

$$A_k - E = \frac{1}{F_k} + \sum_{j=1}^{\infty} U_{k+1}^{(j)} F_{k+j} L_k^{(j)}, \quad k = 0,1,\ldots. \tag{2.13}$$

In the abbreviated notation, we may also write Eq. (2.12) as a recurrent definition of U's and L's,

$$U_{k+1}^{(l)} = C_{k+1}^{(l)} - \mathbf{U}_{k+1}^{T[l]} \cdot \widetilde{F}_{k+l+1} \cdot \mathbf{L}_{k+l},$$
$$L_k^{(l)} = B_k^{(l)} - \mathbf{U}_{k+l+1}^T \cdot \widetilde{F}_{k+l+1} \cdot \mathbf{L}_k^{[l]}, \quad l = \ldots,2,1. \tag{2.14}$$

Then, Eq. (2.13) will represent just a vectorial generalization of the continued fractions (A5),

$$1/F_k = A_k - E - \mathbf{U}_{k+1}^T \cdot \widetilde{F}_{k+1} \cdot \mathbf{L}_k, \quad k = \ldots,2,1,0. \tag{2.15}$$

Thus, in analogy with (A9), the eigenvalue condition $\det(H - E) = 0$ acquires the simple VCF form $1/F_0 = 0$, i.e.,

$$E = A_0 - \mathbf{U}_1^T \cdot \widetilde{F}_1 \cdot \mathbf{L}_0. \tag{2.16}$$

In the original Schrödinger equation, the factorized operator $H - E$ or $H^{\text{eff}} - E$ may be divided by its regular QUQ part. This leads to the generalized Eq. (A10),

$$\begin{pmatrix} 1/F_0 & & \\ L_0^{(1)} & 1/F_1 & \\ L_0^{(2)} & L_1^{(1)} & 1/F_2 \\ & & \ddots \end{pmatrix} \begin{pmatrix} \langle 0|\psi\rangle \\ \langle 1|\psi\rangle \\ \langle 2|\psi\rangle \\ \vdots \end{pmatrix} = 0 \tag{2.17}$$

with the numbers B_k replaced by the VCF vectors $\mathbf{L}_k$. The triangularity of the new matrix simplifies also the wave functions in a way paralleling Eq. (A.11),

$$\langle k|\psi\rangle = (-1)^k \cdot \langle 0|\psi\rangle \cdot \det \begin{pmatrix} F_1 \cdot L_0^{(1)} & 1 & 0 & \cdots & 0 \\ F_2 \cdot L_0^{(2)} & F_2 \cdot L_1^{(1)} & 1 & \cdots & 0 \\ \vdots & & & & \\ F_k \cdot L_0^{(k)} & F_k \cdot L_1^{(k-1)} & & \cdots & F_k \cdot L_{k-1}^{(1)} \end{pmatrix}, \quad k = 1,2,\ldots. \tag{2.18}$$

The VCF construction of bound states is completed.

III. AN ALGEBRAIC TRACTABILITY OF THE SPECIAL, ASYMPTOTICALLY SMOOTH MATRICES H

On an arbitrary fixed level of precision, we may always use truncations of the type

$$\begin{aligned}&\mathbf{B}_i^T \approx (B_i^{(1)},...,B_i^{(t_i)},0,0,...),\\ &\mathbf{C}_j^T \approx (C_j^{(1)},...,C_j^{(s_j)},0,0,...),\end{aligned} \tag{3.1}$$

since the scalar products $\mathbf{C}_i^T \cdot \mathbf{B}_j$ (contributing to the matrix elements of H^2) must converge. Moreover, in the spirit of Appendix B, we shall assume that

$$A_k \approx A_{k+1}, \quad \mathbf{B}_k = \mathbf{C}^*_{k+1} \approx \mathbf{B}_{k+1} = \mathbf{C}^*_{k+2}, \quad k \gg k_0, \tag{3.2}$$

i.e., $t_i = s_i = t \gg 1$ in (3.1). These restrictions still specify a sufficiently broad class of Hamiltonians.

A. The elementary factors

Let us recall (2.11) and interpret each factor as a product of t two-diagonal matrices. For the sake of clarity, we may also write

$$\begin{aligned}&H - E = L^+ D L,\\ &L = L_1 D_1 L_2 D_2 \cdots L_t D_t, \quad t \leqslant \infty,\end{aligned} \tag{3.3}$$

where $D = D^+$ and D_i are diagonal, $(L_i)_{mm} = 1$, $(L_i)_{m+1\,m} \neq 0$, and $(L_i)_{mn} = 0$ otherwise, $i = 1,2,...,t$. The condition (3.2) may be combined with (3.3) in various ways. Here, we shall postulate that L_i are asymptotically constant and write

$$\begin{aligned}Q(H-E)Q \approx QD_t^* \cdot &\begin{pmatrix} 1 & d_t^* & \\ & 1 & d_t^* \\ & & \ddots \end{pmatrix} \cdot D^*_{t-1} \cdot \begin{pmatrix} 1 & d^*_{t-1} & \\ & 1 & d^*_{t-1} \\ & & \cdots \end{pmatrix} \cdot D^*_{t-2} \times \cdots \\ &\times \begin{pmatrix} 1 & d_1^* & \\ & 1 & d_1^* \\ & & \ddots \end{pmatrix} \cdot D \cdot \begin{pmatrix} 1 & & \\ d_1 & 1 & \\ & d_1 & 1 \\ & & & \ddots \end{pmatrix} \cdot D_1 \times \cdots \times \begin{pmatrix} 1 & & \\ d_t & 1 & \\ & d_t & 1 \\ & & & \ddots \end{pmatrix} \cdot D_t \cdot Q.\end{aligned} \tag{3.4}$$

After a partial ordering of the (complex) parameters d_i,

$$|d_i| \geqslant |d_{i+1}|,$$

we may also set $d_i = 0$ beyond some fixed and finite index $t = i_{\max}$ (on any predetermined level of approximation). Moreover, our "far-off-diagonal" truncation may be combined also with another approximation: $QD_1 = \text{const},...,QD_{t-1} = \text{const}$ and, possibly, $QD = \text{const}$. In such a case, we obtain a new representation of the operator $Q(H-E)Q$,

$$Q(H-E)Q \approx QD_t^* \begin{pmatrix} e^{-\alpha_t^*} & e^{\alpha_t^*} & \\ & e^{-\alpha_t^*} & e^{\alpha_t^*} \\ & & \ddots \end{pmatrix} \begin{pmatrix} e^{-\alpha^*_{t-1}} & e^{\alpha^*_{t-1}} & \\ & e^{-\alpha^*_{t-1}} & e^{\alpha^*_{t-1}} \\ & & \ddots \end{pmatrix} \times \cdots \times \begin{pmatrix} e^{-\alpha_t} & & \\ e^{\alpha_t} & e^{-\alpha_t} & \\ & e^{\alpha_t} & e^{-\alpha_t} \\ & & & \ddots \end{pmatrix} D_t Q \tag{3.5}$$

after a reparametrization of $d_i = \exp 2\alpha_i$. Here, all the dependence on indices is to be carried by the diagonal outer factors D_t and D_t^+.

In the simplest $t = 1$ special case, we get the quasiconstant behavior of the Hamiltonian in its asymptotic tridiagonal part,

$$Q(H-E)Q \approx QD_t^* \begin{pmatrix} \cdots \\ \cdots & 0 & e^{2ilm\alpha} & 2\,\text{ch}\,2\,\text{Re}\,\alpha & e^{-2ilm\alpha} & 0 & \cdots \\ \cdots \end{pmatrix} D_t Q. \tag{3.6}$$

In particular, we get the real and symmetric Hamiltonians

$$D_t^{*-1} Q(H-E)QD_t^{-1} \approx \begin{pmatrix} \cdots \\ \cdots & 1 & 2\,\text{ch}\,2\alpha & 1 & \cdots \\ & & \cdots \end{pmatrix} = \begin{pmatrix} \cdots \\ \cdots & 121 & \cdots \\ & \cdots \end{pmatrix} + 4\,\text{sh}^2\,\alpha \cdot I \tag{3.7}$$

for the real α's. Similarly, we get the real asymptotics with $t = 2$,

$$H_{mn} - E\delta_{mn} \approx (D_2^*)_{mm} \left[\binom{4}{2+m-n} + 4(\text{sh}^2\,\alpha_1 + \text{sh}^2\,\alpha_2) \binom{2}{1+m-n} + 16\,\text{sh}^2\,\alpha_1\,\text{sh}^2\,\alpha_2 \delta_{mn} \right] (D_2)_{nn}, \quad m,n \geqslant M \gg 1, \tag{3.8}$$

etc. The general formulas may be found elsewhere.[14]

The ambiguity or symmetry between α_i and $-\alpha_i$ (or $\alpha_i \to -\alpha_i^*$, in general) will be removed later; in brief, it will be shown to intertwine the physical and unphysical wave functions. Thus, in the physical case, we obtain the unique prescriptions from Eq. (2.6). In particular, with $t = 1$, we get the effective matrix element

$$H^{\text{eff}}_{MM} - E \approx (D_1)^2_{MM} \exp(-2\alpha), \quad M \gg 1 \tag{3.9a}$$

from (3.7) or, in general,

$$H^{\text{eff}}_{MM} - E \approx |D_1|^2_{MM} \exp(-2\,\text{Re}\,\alpha), \quad M \gg 1 \tag{3.9b}$$

from (3.6). Similarly, with $t = 2$, we get

$$\begin{aligned} D_2^{-1}(H^{\text{eff}} - E)D_2^{-1} &\approx PD_2^{-1}(H-E)D_2^{-1}P \\ &\quad - \begin{pmatrix} 0 & \cdots & 0 \\ & \cdots & \\ 0\cdots 0, & \exp 2(\alpha_1+\alpha_2), & 2\exp(\alpha_1+\alpha_2)\text{ch}(\alpha_1-\alpha_2) \\ 0\cdots 0, & 2\exp(\alpha_1+\alpha_2)\text{ch}(\alpha_1-\alpha_2), & 4\,\text{ch}^2(\alpha_1-\alpha_2)+\exp 2(\alpha_1+\alpha_2) \end{pmatrix} \\ &\approx \begin{pmatrix} & \cdots & \\ \cdots & 4\,\text{ch}^2(\alpha_1-\alpha_2)+\exp[-2(\alpha_1+\alpha_2)], & 2\exp[-(\alpha_1+\alpha_2)]\text{ch}(\alpha_1-\alpha_2) \\ \cdots & 2\exp[-(\alpha_1+\alpha_2)]\text{ch}(\alpha_1-\alpha_2), & \exp[-2(\alpha_1+\alpha_2)] \end{pmatrix} \end{aligned} \tag{3.10}$$

from (3.8), etc.

B. The difference-equation technique

The Q-projected part of the Schrödinger equation (2.1) may be rewritten in the form factorized in accord with Eq. (2.5),

$$QUQQLQ|\psi\rangle + QUQQLP|\psi\rangle = 0. \tag{3.11}$$

Obviously, the regular factor QUQ is redundant here. Moreover, we may ignore the first t rows ($t < \infty$) and write

$$\tilde{Q}LQ|\psi\rangle = 0, \quad \tilde{Q} = \sum_{n=M+t+1}^{\infty} |n\rangle\langle n|, \tag{3.12}$$

since $\tilde{Q}LP \equiv 0$. Now, the $(t+1)$-term recurrences (3.12) are to be treated as a difference equation of the tth order.[15]

A complete factorization (3.3) of L converts Eq. (3.12) into the asymptotic relations

$$\tilde{Q}DL_1D_1\cdots L_tD_tQ\begin{pmatrix} \langle M+1|\psi\rangle \\ \langle M+2|\psi\rangle \\ \vdots \end{pmatrix} = 0. \tag{3.13}$$

After the simplification (3.5), the latter relations become solvable in a closed form,

$$\langle M|\psi\rangle = \sum_{i=1}^{t} \lambda_i \psi_m^{[i]}, \quad \psi_m^{[i]} \approx (-d_i)^m, \quad m > M \gg 1. \tag{3.14}$$

For all H with t different values of d_i such that $|d_i| \neq 1$, the t independent solutions $\psi_m^{[i]}$ of our asymptotic Schrödinger equation (3.12) contain still the above-mentioned $d_i \to 1/d_i$ ambiguity. The standard requirement of existence of $\psi(r)$ with the finite norm,

$$\sum_{m=0}^{\infty} \langle\psi|m\rangle\langle m|\psi\rangle < \infty \tag{3.15}$$

implies that we may remove the ambiguity via the sufficient condition of convergence in Eq. (3.15),

$$|d_i| < 1, \quad i = 1,2,\ldots,t. \tag{3.16}$$

The boundary of the physical region with some $|d_i| = 1$ must be investigated separately.

In the above discussion, our use of factorizations is redundant. The kth row of the band-matrix Schrödinger equation

$$\begin{aligned} &B^{(t)}_{k-t}\langle k-t|\psi\rangle + B^{(t-1)}_{k-t+1}\langle k-t+1|\psi\rangle \\ &\quad + \cdots + C^{(t)}_{k+1}\langle k+t|\psi\rangle = 0 \end{aligned} \tag{3.17}$$

may be directly understood as a difference equation of the $2t$th order.[15] Its $2t$ independent solutions $\psi_k^{[i]}$, $i = \pm 1, \pm 2,\ldots, \pm t$ complement the set (3.14) simply via the replacement $d_i \to 1/d_i$ for the negative super/subscripts i. We may preserve the former notation and denote the Jost solutions[16] [compatible with (3.15)] by the upper positive superscripts.

The physical parameters λ_i and E in (3.14) are to be determined from the first $M+t+1$ rows of the Schrödinger equation omitted in our asymptotic subsystem of Eqs. (3.12). Vice versa, a variation of the submatrix PHP leads to a change of these coefficients in general. Hence the last t rows of the effective Eq. (2.3),

$$\begin{aligned} &B^{(t)}_{M-2t+j}\psi_{M-2t+j} + \cdots + B^{(j)}_{M-t}\psi_{M-t} \\ &\quad + g_{j1}\psi_{M-t+1} + \cdots + g_{jt}\psi_M = 0, \quad j = 1,2,\ldots,t, \end{aligned} \tag{3.18}$$

must be satisfied by each of the t independent Jost states $\psi_k^{[i]}$, $k > M-2t$, $i = 1,2,\ldots,t$. In a compact notation with the $(t\times t)$-dimensional matrices $g = g_m$, $(b_{m-1})_{ij} = B^{(t+i-j)}_{M-2t+j}$ ($M+1 = t_0 + mt$, cf. Appendix C), and

$$x_{ij} = \psi^{[j]}_{M-2t+i}, \quad y_{ij} = \psi^{[j]}_{M-t+i}, \quad i,j = 1,2,\ldots,t,$$

we must satisfy, therefore, the matrix identity $b_{m-1}x + g_my = 0$, i.e.,

$$g_m = -b_{m-1}xy^{-1}. \tag{3.19}$$

This formula is very important: Our knowledge of a complete system of Jost solutions becomes equivalent to an explicit knowledge of the effective Hamiltonian H^{eff}. Equation (3.19) complements the previous MCF or VCF definition of g_m and is in fact the most natural generalization of Eq. (2.7) to all $t \gg 1$.

IV. SCHRÖDINGER EQUATION WITH POLYNOMIAL POTENTIALS

A. The semirelativistic $p-r$ symmetrization of the Hamiltonian

A semirelativistic form or extension[10]

$$(T+V)\psi=E\psi,\quad V=\sum_{i=0}^{m_1} g_i r^{2i+2},$$
$$T=\sum_{j=0}^{m_2} h_i p^{2i+2}=(c^4+p^2c^2)^{1/2}-c^2+O(p/c)^{2m_2+4} \tag{4.1}$$

of the anharmonic oscillator equation (1.1) degenerates to the exactly solvable harmonic oscillator for $m_1=m_2=0$. With $m_1=m_2=2q$, it remains symmetric with respect to the Fourier transformation $p\leftrightarrow r$. Here, it may be used as an illustrative example of the asymptotically band-matrix Hamiltonian since, in accord with Ref. 10, the asymptotically dominant part of H may be given the form

$$(T+V-E)_{mn}\approx\binom{4q+2}{2q+1+2m-2n}m^{2q+1}$$
$$\times\left(1+O\left(\frac{1}{m}\right)\right),\quad m\gg 1. \tag{4.2}$$

For the sufficiently large indices m and n, we may write

$$(T+V-E)_{mn}\approx\binom{2q}{q+m-n}$$
$$+\sum_{l=1}^{q}\rho_l^{(q)}\binom{2q-2l}{q-l+m-n}, \tag{4.3}$$

where

$$\rho_1^{(1)}=\tfrac{4}{3},\quad \rho_1^{(2)}=8,\quad \rho_2^{(2)}=16/5,$$
$$\rho_1^{(3)}=20,\quad \rho_2^{(3)}=48,\quad \rho_3^{(3)}=64/7, \tag{4.4}$$

etc.

In the first nontrivial example with $q=1$ and recurrences (4.1),

$$\psi_{n-1}+\tfrac{10}{3}\psi_n+\psi_{n+1}=0,\quad n\gg 1 \tag{4.5}$$

we may put $\psi_n^{[t1]}\sim(-d)^n$ and obtain $d=\frac{1}{3}$ or $d=3$. In this case, the auxiliary VCF quantities degenerate to the ordinary continued fractions, and the finite approximants with $F_{M+1}=0$ may be given an explicit form,

$$F_{M+1-k}=d\,[(d^{2k}-1)/(d^{2k+2}-1)],\quad k=0,1,\ldots. \tag{4.6}$$

This converges to the value $d_1=\min(d,d^{-1})$ that is always smaller than 1. We observe a compatibility of the variational truncation with our difference-equation requirement (3.16).

In a straightforward way, the same analysis may be repeated for $t=q>1$ as well. For example, the $q=2$ recurrences (4.1),

$$\psi_{n-2}+12\psi_{n-1}+25.2\psi_n+12\psi_{n+1}$$
$$+\psi_{n+2}=0,\quad n\gg 1, \tag{4.7}$$

may be solved by the ansatz (3.14) which gives[6]

$$d_1=1/d_{-1}=5d_2=5/d_{-2}=5-2\sqrt{5}\approx 0.527 \tag{4.8}$$

in an elementary way.

For a general value of $q=t$, the VCF convergence becomes less transparent. A basic idea of its analysis[16] may be visualized as follows. First, we recall the partitioned notation of Appendix C and definitions

$$a_k-E=h_k s_k+l_{k+1}u_{k+1},\quad b_k=h_{k+1}u_{k+1},$$
$$c_{k+1}=l_{k+1}s_{k+1}$$

of the $(t\times t)$-dimensional matrices entering the VCF/MCF mapping $g_{k+1}(=h_{k+1}s_{k+1})\rightarrow g_k$ (C5). By means of an ansatz

$$g_k=h_k s_k+l_{k+1}p_k u_{k+1}, \tag{4.9}$$

we convert these recurrences into an equivalent prescription

$$p_k=\gamma_1 p_{k+1}/(1+\rho p_{k+1})\gamma_2,$$
$$\gamma_1=h_{k+1}^{-1}l_{k+2},\quad \gamma_2=u_{k+2}/s_{k+1},\quad \rho=\gamma_2\gamma_1. \tag{4.10}$$

Now, assuming that the k dependence of all the matrices is sufficiently weak here, the VCF/MCF convergence may be reinterpreted as a convergence of iterations of the mapping (4.10) performed at a fixed index $k\gg 1$. We may employ a spectral representation of the auxiliary matrices

$$\gamma_1=\sum_{\epsilon_1\in\Theta_1}|\epsilon_1\rangle\epsilon_1\langle\epsilon_1|,\quad \gamma_2=\sum_{\epsilon_2\in\Theta_2}|\epsilon_2\rangle\epsilon_2\langle\epsilon_2|, \tag{4.11}$$

and denote their maximal eigenvalues by a zero superscript. Immediately, we may notice then that the repeated multiplication of p's by γ's in (4.10) suppresses all the components not corresponding to $\epsilon_i=\epsilon_i^{(0)}$, $i=1,2$.

For the simplest case with the nondegenerate spectra Θ_i and unique values of $\epsilon_i^{(0)}$, $i=1,2$, a sufficient number of iterations in (4.10) converts the dominant component of p_k into a separable expression,

$$p_k=|\epsilon_1^{(0)}\rangle\varkappa_k\langle\epsilon_2^{(0)}|+\text{corrections}. \tag{4.12}$$

Vice versa, an insertion of (4.12) changes the VCF/MCF mapping $g_{k+1}\rightarrow g_k$ or $p_{k+1}\rightarrow p_k$ into an approximatively one-dimensional mapping $\varkappa_{k+1}\rightarrow\varkappa_k$,

$$\varkappa_k=\epsilon_1^{(0)}\epsilon_2^{(0)}\varkappa_{k+1}/(1+\epsilon_1^{(0)}\epsilon_2^{(0)}\langle\epsilon_2^{(0)}|\epsilon_1^{(0)}\rangle\varkappa_{k+1}) \tag{4.13}$$

with an easy analysis of convergence.

B. Effective Hamiltonians in the nonrelativistic case

The nonrelativistic $t=2$ analog (1.1) of Eq. (4.7),

$$\psi_{n-2}+4\psi_{n-1}+6\psi_n+4\psi_{n+1}+\psi_{n+2}=0 \tag{4.14}$$

leads immediately to a quadruple leading-order asymptotic degeneracy[7] $d_i=1$, $i=1$–4 of the Jost solutions (3.14). This result is easily generalized to any t. Indeed, from the initial values

$$A_{M+1}=A_{M+2}=\cdots=0,$$
$$\mathbf{B}_M=\mathbf{B}_{M+1}=\cdots=\mathbf{C}_{M+1}=\cdots=0, \tag{4.15}$$
$$\mathbf{B}_{M-1}^{[1]}=\mathbf{B}_{M-2}^{[2]}=\cdots=\mathbf{C}_M^{[1]}=\cdots=0,\quad M\gg 1$$

we obtain, after a reasonable amount of the algebraic manipulations, the sequences

$$F_{M+1-k}=\frac{k(k+1)\cdots(k+t-1)(1+O(k/M))}{(k+t)(k+t+1)\cdots(k+2t-1)},$$

$$U^{(j)}_{M+1-k}$$
$$=\binom{t}{j}\frac{(k+t+j)(k+t+j+1)\cdots(k+2t-1)}{(k+j)(k+j+1)\cdots(k+t-1)}$$
$$\times\left(1+O\left(\frac{k}{M}\right)\right),\quad j=1,2,\ldots,t,\quad k=1,2,\ldots, \tag{4.16}$$

i.e., a rigorous solution of the VCF recurrences (2.14) and (2.15) in the asymptotic region.

In contrast to the $p\leftrightarrow r$ symmetric equation (4.1) with

$$|\langle n+1|\psi\rangle/\langle n|\psi\rangle|\approx\max_{|i|}|d_i|<1,\quad n\gg1, \tag{4.17}$$

our present degeneracy of asymptotics need not be considered as a shortcoming. In fact, we may write

$$\langle n|\psi\rangle=(-1)^n\exp\xi_n, \tag{4.18}$$

where ξ_n should be some asymptotically smooth function of the index n. Its important merit lies in a possibility of inserting the Taylor series

$$\exp\xi_{n+k}=\exp\xi_n+k\frac{d}{dn}\exp\xi_n+\cdots$$
$$=\exp\xi_n(1+k\xi'_n+\tfrac{1}{2}k^2(\xi''_n+\xi'^2_n)+\cdots) \tag{4.19}$$

in the left-hand side $V|\psi\rangle$ of equations of the type (4.14), $QV|\psi\rangle=0$. Then, whenever we replace the right-hand side zero by the leading-order estimate of the kinetic energy contribution $-T|\psi\rangle\approx4\psi_n$, $n\gg1$, we obtain an estimate of ξ_n or

$$\xi'_n\approx\text{const}\times n^{(1-t)/2t},\quad n\gg1, \tag{4.20}$$

in accord with Ref. 11.

In the next step, we may notice that all the corrections in (4.19) are given as derivatives of ξ_n. An inclusion of the precise matrix elements of H may only modify and convert Eq. (4.20) into a series

$$\xi'_n=\sum_{m=1}^{L}\beta_m\rho^m+O(\rho^{L+1}),\quad\rho=n^{-1/2t}\ll1, \tag{4.21}$$

where, in accord with Eq. (4.20),

$$\beta_1=\cdots=\beta_{t-2}=0,\quad\beta_{t-1}\neq0. \tag{4.22}$$

In analogy with the Jost solutions pertaining to the Hamiltonians with the nondegenerate parametrization (3.5), our present expansion (4.21) may also be used as an ansatz. It transforms the asymptotic anharmonic oscillator Schrödinger equation [say, Eq. (3.17) with $k=n\gg1$] into a power-series requirement of the implicit general form

$$\sum_{k=0}^{\infty}\rho^kR_k(E,l,g_1,\ldots,g_t,\beta_1,\beta_2,\ldots)=0. \tag{4.23}$$

Due to a linear independence of the powers of ρ, the conditions

$$R_k=0,\quad k=0,1,\ldots, \tag{4.24}$$

are to be solved as a nonlinear algebraic set of coupled equations for the coefficients β_i in (4.21) or (4.18) and (4.19).

1. t=2

Let us start our analysis from the simplest nontrivial quartic oscillator problem (1.1) with $t=2$. Performing the algebraic construction of the functions R_k in (4.24) on the computer (in REDUCE), we find that the first four items are identically zeros. This reflects the asymptotically correct behavior of our ansatz (4.18). From $R_4=0$ we obtain

$$\beta_1^4=-4/g_2. \tag{4.25}$$

In full analogy with our preceding discussion, the four independent complex roots $\beta_1^{(j)}$ of this equation with $j=\pm1,\pm2$ comprise also the physical Jost coefficients

$$\beta_1^{(1,2)}=(-1\pm i)\cdot|g_2^{-1/4}|. \tag{4.26}$$

Similarly, the further items of Eq. (4.24) remain linear in the unknown coefficients and determine the respective sequence of their values

$$\beta_2=0,$$
$$\beta_3=\tfrac{1}{8}(1\pm i)(\tfrac{1}{3}-g_1)|g_2^{-3/4}|,$$
$$\beta_4=-\tfrac{5}{8},$$
$$\beta_5=\tfrac{1}{16}(1\mp i)[(E+2l+3)g_2 \tag{4.27}$$
$$+\tfrac{1}{8}(g_1+1)^2-\tfrac{1}{5}]\cdot|g_2^{-5/4}|,$$
$$\beta_6=(\mp i/32)(3-g_1)\cdot|g_2^{-1/2}|,$$
$$\vdots\,.$$

Up to a misprint in β_5 and missing β_6, they agree with the $g_2=1$ results of Ref. 11, i.e., with the wave function asymptotics

$$\psi_n^{[j]}=(-1)^nn^{-5/8}\exp(\tfrac{4}{3}\beta_1^{(j)}n^{3/4}+4\beta_3^{(j)}n^{1/4})$$
$$\times\exp(-4\beta_5^{(j)}n^{-1/4}-2\beta_6^{(j)}n^{-1/2}$$
$$+O(n^{-3/4})),\quad j=1,2, \tag{4.28}$$

tested and verified numerically for $g_1=1$. Of course, such a choice simplifies also the higher-order coefficients here,

$$\beta_6^{(j)}=\tfrac{1}{32}\beta_1^{(j)2},$$
$$\beta_7^{(j)}=\tfrac{1}{32}\beta_1^{(j)3}(-4\alpha^2+5\alpha+\tfrac{1}{4}E+\tfrac{101}{112}),\quad\alpha=\tfrac{1}{2}l+\tfrac{3}{4},$$
$$\beta_8=\tfrac{5}{8}\alpha-\tfrac{3}{32}E+\tfrac{1}{32},$$
$$\beta_9^{(j)}=\tfrac{1}{32}\beta_1^{(j)}(11\alpha^2+\tfrac{1}{4}\alpha(10E-21) \tag{4.29}$$
$$-\tfrac{1}{16}(3E^2-3E+\tfrac{835}{36})),$$
$$\beta_{10}^{(j)}=\tfrac{1}{32}\beta_1^{(j)2}(-3\alpha^2+\tfrac{3}{2}\alpha-\tfrac{3}{8}E+\tfrac{23}{64}),$$
$$\vdots\,.$$

2. t=3

In the sextic anharmonic oscillator example (1.1) with $g_1=1$ (scaling of the scale) and $t=3$, we may proceed along the same lines as above. After confirming Eq. (4.22) as a consequence of (4.24) at $k\leqslant6$, we obtain the first nontrivial algebraic equation $R_k=0$ at $k=12$,

$$\beta_2^{(j)6}=4/g_3. \tag{4.30}$$

We may specify the Jost roots $\beta_2^{(j)}$, $j=1$, 2, and 3, by the simple condition

$$\text{sgn Re}\,\beta_{t-1}^{(j)}=-1,\quad j=1,2,\ldots,t. \tag{4.31}$$

In the forthcoming steps, we obtain the zero odd coefficients

$$\beta_3=\beta_5=\beta_7=\beta_9=0,$$

and the nontrivial contributions

$$\begin{aligned}&\beta_4^{(j)}=\tfrac{1}{24}g_2\beta_2^{(j)5},\\&\beta_6^{(j)}=\tfrac{1}{24}[-16+(-1+g_2^2/4g_3)\beta_2^{(j)3}],\\&\beta_8^{(j)}=\frac{1}{24}\Big[\frac{1}{6}g_2\beta_2^{(j)4}\\&\qquad+\Big(\frac{7}{216}\frac{g_2^3}{g_3^2}-\frac{1}{2}\frac{g_2}{g_3}-8\alpha-E\Big)\beta_2^{(j)}\Big],\end{aligned}\tag{4.32}$$

to the physical wave functions

$$\begin{aligned}\langle n|\psi\rangle=&\sum_{j=1}^{3}(-1)^n\lambda_j n^{\beta_6^{(j)}}\\&\times\exp\Big(\frac{3}{2}\beta_2^{(j)}n^{2/3}+\frac{g_2}{2\beta_2^{(j)}g_3}n^{1/3}\Big)\\&\times\exp(-3\beta_8^{(j)}n^{-1/3}+O(n^{-2/3})),\quad n\gg 1.\end{aligned}\tag{4.33}$$

3. t=4

A knowledge of the first few solutions with increasing t simplifies the manipulations in REDUCE—we may incorporate there immediately the relations (4.22), formula

$$\beta_{t-1}^{2t}=(-1)^{t+1}\cdot(4/g_t)\tag{4.34}$$

derived in the second paper of Ref. 11 and further hypotheses inspired by the similarities of the preceding solutions. We must be careful, of course. In particular, the octic oscillator analog

$$\begin{aligned}\psi_n^{[j]}=&(-1)^n n^{\beta_8^{(j)}}\exp(\tfrac{8}{5}\beta_3^{(j)}n^{5/8}+\tfrac{8}{3}\beta_5^{(j)}n^{3/8}\\&+8\beta_7^{(j)}n^{1/8}\exp(-8\beta_9^{(j)}n^{-1/8}+O(n^{-1/4})),\\&n\gg 1,\quad t=4,\end{aligned}\tag{4.35}$$

of (4.33) or (4.28) shows that an assumption $\beta_8=0$ is not correct. In detail, Eq. (4.24) with $k=24, 26, 28, 29, 30$, etc. leads to the respective sets of coefficients

$$\begin{aligned}&\beta_3^{(j)}=(-4/g_4)^{1/8},\\&\beta_5^{(j)}=-\tfrac{1}{32}g_3\beta_3^{(j)7},\\&\beta_7^{(j)}=\tfrac{1}{32}\beta_3^{(j)5}\Big(g_2-\frac{5}{16}\frac{g_3^2}{g_4}\Big),\quad\beta_8^{(j)}=-\frac{11}{16},\\&\beta_9^{(j)}=\frac{1}{32}\beta_3^{(j)3}\Big(-\frac{4}{3}+\frac{3g_2g_3}{8g_4}-\frac{11g_3^3}{128g_4^2}\Big),\\&\quad\vdots\end{aligned}\tag{4.36}$$

with the scaled coupling $g_1=1$ again. These formulas are very suitable for an analysis of the various kinds of limits, but this is beyond the scope of the present paper.

V. SUMMARY

One of the most efficient numerical approaches to polynomial interactions is known to be a diagonalization in the standard oscillator basis. In a quasiperturbative, less numerical setting, this technique has recently been shown equivalent to a systematic algebraic construction of the MCF fixed-point approximants. Here, we have succeeded in solving the corresponding nonlinear systems of algebraic equations in a non-numerical manner. Up to the octic anharmonic oscillators, the explicit asymptotic-series representations of the effective Hamiltonians have been obtained.

Methodically, the two aspects of the technique deserve special attention. First, a similarity in structure of the asymptotically degenerate and nondegenerate Hamiltonians has been recovered. Conceptually, this makes the whole method extremely simple. The present constructions illustrate also its purely pragmatic efficiency.

Second, the underlying factorization has been given a "final" form—its vectorially partitioned character does not necessitate any sparse-matrix assumption anymore. In this sense, we believe in its further methodical development, especially via the various straightforward extensions of the underlying ansatz.

APPENDIX A: TRIDIAGONAL HAMILTONIANS AS A METHODICAL GUIDE

At the very beginning of the standard Lanczos numerical tridiagonalization of H, we have to pick up an arbitrary trial vector $|0\rangle$. Then, by means of its repeated multiplication by the operator H, we may generate the basis,[12]

$$\begin{aligned}&|1\rangle=(1/B_0)[H|0\rangle-|0\rangle\cdot A_0],\\&\quad\vdots\\&|k+1\rangle=(1/B_k)[H|k\rangle-|k\rangle\cdot A_k\\&\qquad-|k-1\rangle C_k],\quad k=1,2,\ldots.\end{aligned}\tag{A1}$$

The natural condition of orthonormality

$$\langle m|n\rangle=\begin{cases}0, & m\neq n,\\1, & m=n\geqslant 0,\end{cases}\tag{A2}$$

determines all the coefficients in (A1) uniquely.

The operator H becomes represented by the tridiagonal matrix

$$H=\begin{pmatrix}A_0 & C_1 & & \\ B_0 & A_1 & C_2 & \\ & B_1 & A_2 & C_3\\ & & & \ddots\end{pmatrix}\tag{A3}$$

in the basis (A1). As a consequence, we may introduce the factorization (2.5),

$$H-E=\begin{pmatrix}1 & C_1F_1 & \\ & 1 & C_2F_2\\ & & \ddots\end{pmatrix}\cdot\begin{pmatrix}1/F_0 & & \\ & 1/F_1 & \\ & & \ddots\end{pmatrix}\cdot\begin{pmatrix}1 & & \\ F_1B_0 & 1 & \\ & F_2B_1 & 1\\ & & \ddots\end{pmatrix},\tag{A4}$$

where the auxiliary quantities F_k have only to satisfy the recurrences

$$1/F_k = A_k - E - C_{k+1}F_{k+1}B_k, \quad k \geqslant 0. \tag{A5}$$

Formally, this enables us to rewrite the secular determinant in the factorized form

$$\det(H-E) = \left(\prod_{k=0}^{\infty} F_k\right)^{-1}. \tag{A6}$$

In the computational practice,[17] we employ usually the truncation of H,

$$\begin{aligned} &A_{N+1} = A_{N+2} = \cdots = 0, \\ &C_{N+1} = \cdots = 0, \quad B_N = \cdots = 0, \quad N \ll 1. \end{aligned} \tag{A7}$$

In the limit $N \to \infty$, this leads to the results equivalent to an exact solution. In the present setting, this simplifies also an interpretation of recurrences (A5)—we may use $F_{N+1} = 0$ as an initial value and identify the quantities F_k with the analytic continued fractions,[18]

$$F_k = (A_k - E - C_{k+1}B_k/(A_{k+1} - E - \cdots))^{-1}. \tag{A8}$$

For an arbirtrary finite cutoff parameter $N < \infty$, the Schrödinger equation (1.1) possesses a nontrivial solution if and only if $\det(H-E) = 0$. Here, we may compare Eqs. (A5) and (A6) and see that the binding energies will coincide with the roots of the analytic continued fractional "Green's function" $1/F_0$,

$$E = A_0 - C_1F_1(E)B_0. \tag{A9}$$

Moreover, we may also omit the regular factor from Eq. (1.1) and, noticing that the first row of the resulting new form of our Schrödinger equation

$$\begin{pmatrix} 1/F_0 & \cdots & & \\ B_0 & 1/F_1 & \cdots & \\ 0 & B_1 & 1/F_2 & \\ & & \ddots & \end{pmatrix} \begin{pmatrix} \langle 0|\psi\rangle \\ \langle 1|\psi\rangle \\ \langle 2|\psi\rangle \\ \cdots \end{pmatrix} = 0 \tag{A10}$$

becomes satisfied identically [cf. Eq. (A9)], we obtain an explicit continued fractional formula

$$\begin{aligned} \langle k|\psi\rangle &= -F_kB_{k-1}\langle k-1|\psi\rangle \\ &= (-1)^k \prod_{m=0}^{k-1} B_mF_{m+1}\cdot\langle 0|\psi\rangle. \end{aligned} \tag{A11}$$

This defines the bound states.

APPENDIX B: THE FIXED POINT EXPANSIONS

In practice, a continued-fractional convergence $N \to \infty$ is usually very quick. This may easily be understood as a consequence of the weakening k dependence of the mapping $F_{k+1} \to F_k$ (A5) for the increasing indices $k \gg 1$. Under such an assumption (summarizing in fact just the numerical experience), our knowledge of the large number of quantities $F_N, F_{N-1}, \ldots, F_{k+1}, F_k$ becomes redundant. Indeed, a weak k dependence of F_k's for $k \gg 1$ implies that we may expect that $F_{k+1} \approx F_k$ may be approximated by a "fixed-point" root $F_k^{[0]}$ of the simple quadratic equation

$$1/F_k^{[0]} = A_k - E - C_{k+1}F_k^{[0]}B_k, \quad k \geqslant 0. \tag{B1}$$

An ambiguity of this definition may easily be removed by means of the stability criterion—the physical root of (B1) becomes unique.

In accord with Eq. (A9), the quantity $1/F_0$ (or, in general, $1/F_k$) may be interpreted as a component of the Feshbach effective Hamiltonian.[8,9] In a perturbative spirit of Ref. 7, we may also study the corrections

$$F_k^{(1)} = F_k - F_k^{[0]}. \tag{B2}$$

Indeed, when we rewrite Eq. (A5) in a new form

$$1/(F_k^{(1)} + F_k^{[0]}) = A_k - E - C_{k+1}(F_{k+1}^{(1)} + F_{k+1}^{[0]})B_k, \tag{B3}$$

we may subtract the definition (B11) and eliminate, say, the parameters $A_k - E$,

$$\begin{aligned} &(1/F_k^{[0]})F_k^{(1)}[1/(F_k^{(1)} + F_k^{[0]})] \\ &\quad = C_{k+1}(F_{k+1}^{(1)} + F_{k+1}^{[0]} - F_k^{[0]})B_k. \end{aligned}$$

This may be rewritten as a new rational mapping

$$F_k^{(1)} = (\alpha_k^{(1)} + \beta_k^{(1)}F_{k+1}^{(1)})/(\gamma_k^{(1)} - F_{k+1}^{(1)}), \quad k = 0,1,\ldots \tag{B4}$$

with the k-dependent parameters

$$\begin{aligned} \alpha_k^{(1)} &= F_k^{[0]}(F_{k+1}^{[0]} - F_k^{[0]}), \quad \beta_k^{(1)} = F_k^{[0]}, \\ \gamma_k^{(1)} &= (B_kC_{k+1})^{-1}(A_k - E) - F_{k+1}^{[0]} \\ &= (F_k^{[0]}B_kC_{k+1})^{-1} + F_k^{[0]} - F_{k+1}^{[0]}. \end{aligned} \tag{B5}$$

Obviously, the subtraction of the type (B2) may easily be iterated. First, we have to specify the higher-order fixed-point approximants $F_k^{[n]}$ as the roots of the generalized Eqs. (B1),

$$F_k^{[n]} = (\alpha_k^{(n)} + \beta_k^{(n)}F_k^{[n]})/(\gamma_k^{(n)} - F_k^{[n]}), \quad n \geqslant 1. \tag{B6}$$

Next, we notice a uniqueness of the definition (B6): one of the roots represents just a return to the solution discarded in the zero-order equation (B1). Thus assuming that the iterations

$$F_k = F_k^{[0]} + F_k^{[1]} + \cdots + F_k^{[n]} + F_k^{(n+1)} \tag{B7}$$

converge, the smallest roots of Eq. (B6) may be treated as "physical."

In a way analogous to the derivation of Eq. (B4), we may rewrite Eq. (A5) in the equivalent form

$$\begin{aligned} F_k^{(n+1)} = (\alpha_k^{(n+1)} + \beta_k^{(n+1)}F_{k+1}^{(n+1)})/(\gamma_k^{(n+1)} \\ - F_{k+1}^{(n+1)}), \quad k,n \geqslant 0. \end{aligned} \tag{B8}$$

This completes our fixed-point (FP) construction of the new expansion (B7). Indeed, we obtain the recurrent definitions of the relevant coefficients

$$\begin{aligned} \gamma_k^{(n+1)} &= \gamma_k^{(n)} - F_{k+1}^{[n]} = \frac{A_k - E}{B_kC_{k+1}} - \sum_{m=0}^{n} F_{k+1}^{[m]}, \\ \alpha_k^{(n+1)} &= \alpha_k^{(n)} + \beta_k^{(n)}F_{k+1}^{[n]} - \gamma_k^{(n+1)}F_k^{[n]} \\ &= \sum_{m=0}^{n}\sum_{l=0}^{n} F_k^{[m]}F_{k+1}^{(l)} \\ &\quad + \frac{1}{B_kC_{k+1}}\left[1 - (A_k - E)\sum_{m=0}^{n} F_k^{[m]}\right], \\ \beta_k^{(n)} &= \beta_k^{(n)} + F_k^{[n]} = \sum_{m=0}^{n} F_k^{[m]} \end{aligned} \tag{B9}$$

valid for any fixed FP order n and variable index k. In the large-n limit, we get

$$\beta_k^{(n)} \to F_k, \quad \gamma_k^{(n)} \to (B_k F_k C_{k+1})^{-1}, \quad \alpha_k^{(n)} \to 0. \quad \text{(B10)}$$

These relations enable us to prove or analyze the convergence of the FP alternative (B7) to the continued fractional expansion (A5) in detail. Their generalization to $t > 1$ is straightforward. It may be found elsewhere.[7,16]

APPENDIX C: THE VCF/MCF EQUIVALENCE FOR $t < \infty$

For the (2t + 1)-diagonal matrices

$$H = \begin{pmatrix} A_0 & C_1^{(1)} & \cdots & C_1^{(t)} & 0 & & & & \cdots & 0 \\ & & & \cdots & & & & & & \\ 0 & \cdots & & 0 & B_{k-t}^{(t)} & \cdots & C_{k+1}^{(t)} & 0 & \cdots & 0 \\ & & & \cdots & & & & & & \\ 0 & \cdots & & & 0 & B_{N-t}^{(t)} & B_{N-t+1}^{(t-1)} & \cdots & B_{N-1}^{(1)} & A_N \end{pmatrix}, \quad N \leqslant \infty, \quad \text{(C1)}$$

a block-tridiagonal repartitioning may be introduced,

$$H = \begin{pmatrix} a_0 & c_1 & 0 & & \cdots & 0 \\ b_0 & a_1 & c_2 & 0 & \cdots & 0 \\ & & \cdots & & & \\ 0 & \cdots & 0 & b_{m-1} & & a_n \end{pmatrix}, \quad nt + t_0 = N + 1, \quad \text{(C2)}$$

with the $(t_k \times t_k)$-dimensional submatrices a_k and $t_0 \leqslant t_1 = t_2 = \cdots = t$. Then, the VCF factorization (2.11) may either be partitioned,

$$H - E = \begin{pmatrix} h_0 & l_1 & & \\ & h_1 & l_2 & \\ & & \ddots & \\ & & & h_n \end{pmatrix} \times \begin{pmatrix} s_0 & & & \\ u_1 & s_1 & & \\ & & \ddots & \\ & & u_n & s_n \end{pmatrix}, \quad \dim h_k = t_k, \quad \text{(C3)}$$

or replaced by the new form of the general decomposition (2.5), namely, by the formula

$$H - E = \begin{pmatrix} I & c_1 f_1 & & \\ & I & c_2 f_2 & \\ & & \ddots & \\ & & & I \end{pmatrix} \times \begin{pmatrix} 1/f_0 & & & \\ & 1/f_1 & & \\ & & \ddots & \\ & & & 1/f_n \end{pmatrix} \times \begin{pmatrix} I & & & \\ f_1 b_0 & I & & \\ & & \ddots & \\ & & f_n b_{n-1} & I \end{pmatrix}, \quad n \leqslant \infty. \quad \text{(C4)}$$

The related recurrences

$$1/f_k = a_k - EI - c_{k+1} f_{k+1} b_k, \quad k = n, n-1, \ldots, 0, \quad \text{(C5)}$$

define simply the $(t \times t)$-dimensional generalization of the continued fractions[5] in the limit $N \to \infty$.

A comparison of Eqs. (C3) and (C4) implies that

$$\begin{aligned} f_k &= (h_k s_k)^{-1}, \quad k = n, n-1, \ldots, 0, \\ b_{m-1} &= h_m u_m, \quad c_m = l_m s_m, \quad m = 1, 2, \ldots, n, \end{aligned} \quad \text{(C6)}$$

so that the MCF quantities become defined uniquely in terms of their VCF counterparts. Whenever we require that, e.g., $(s_m)_{ii} = 1$, the opposite is also true for a broad class of Hamiltonians.

[1] S. Flügge, *Practical Quantum Mechanics* (Springer, New York, 1971).

[2] C. Quigg and J. L. Rosner, Phys. Rep. **56**, 167 (1979).

[3] C. Itzykson and J. B. Zuber, *Quantum Field Theory* (McGraw-Hill, New York, 1980).

[4] J. Makarewicz, J. Phys. A **17**, 1461 (1984), and references contained therein.

[5] S. Graffi and V. Grecchi, Lett. Nuovo Cimento **12**, 425 (1975).

[6] M. Znojil, J. Phys. A **16**, 3313 (1983).

[7] M. Znojil, J. Math. Phys. **25**, 2979 (1984).

[8] H. Feshbach, Ann. Phys. (NY) **5**, 357 (1958).

[9] A. H. Wilson, Proc. R. Soc. London Ser. A **118**, 617 (1928).

[10] M. Znojil, Phys. Rev. D **24**, 903 (1981).

[11] M. Znojil, K. Sandler, and M. Tater, J. Phys. A **18**, 2541 (1985); M. Znojil, Phys. Lett. A **114**, 349 (1986).

[12] C. Lanczos, J. Res. Natl. Bur. Stand. **45**, 255 (1950).

[13] A. Duncan and R. Roskies, Phys. Rev. D **32**, 3277 (1985).

[14] M. Znojil, J. Phys. A **16**, 4001 (1983).

[15] N. E. Nörlund, *Vorlesungen u. Differenzenrechnung* (Springer, Kopenhagen, 1923).

[16] M. Znojil, J. Phys. A **17**, 1603, 1611 (1984).

[17] F. Acton, *Numerical Methods that Work* (Harper and Row, New York, 1970).

[18] H. S. Wall, *Theory of Analytic Continued Fractions* (Van Nostrand, New York, 1948).

Exact behavior of Jost functions at low energy

Martin Klaus
Department of Mathematics, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061

(Received 1 July 1987; accepted for publication 31 August 1987)

For Schrödinger operators with central potential $q(r)$ and angular momentum l, the behavior of the Jost function $F_l(k)$ as $k\to 0$ is investigated. It is assumed that $\int_0^\infty dr\,(1+r)^\sigma|q(r)|<\infty$, where $\sigma\geqslant 1$. Situations where q is integrable with $1\leqslant\sigma<2$, but not with $\sigma\geqslant 2$ are of particular interest. For potentials satisfying $q(r)\sim q_0r^{-2-\epsilon}$ $(0<\epsilon\leqslant 1)$ and $l=0$, the leading behavior of $F_0(k)$ and the phase shift $\delta_0(k)$ as $k\to 0$ is derived. Also comments are made on the differentiability properties of the Jost solutions with respect to the variable k at $k=0$. For $\sigma=1$ Levinson's theorem is proved, thereby clarifying some questions raised recently by Newton [J. Math. Phys. 27, 2720 (1986)].

I. INTRODUCTION

In this paper we study the low-energy behavior of Jost functions and phase shifts of the three-dimensional Schrödinger equation with central potential $q(r)\in L^1_\sigma$, where

$$L^1_\sigma=\left\{q\left|\int_0^\infty(1+r)^\sigma|q(r)|dr<\infty,\ \sigma\geqslant 1.\right.\right\}\qquad(1.1)$$

Our main concern are potentials that are in L^1_σ with $1\leqslant\sigma<2$, but not necessarily in L^1_2. We were stimulated by a recent paper of Newton[1] on this subject and in particular by one result which we recall here briefly. Let $F_l(k)$ denote the Jost function corresponding to angular momentum l $(l=0,1,2,\ldots)$ and assume that $q\in L^1_\sigma$ with $1\leqslant\sigma<2$ if $l=0$ or $1\leqslant\sigma<3$ if $l\geqslant 1$. Then Newton proved that

$$F_l(k)=F_l(0)+o(k^{\sigma-1}).\qquad(1.2)$$

So, if $F_l(0)=0$, then it is consistent with (1.2) if

$$F_l(k)=ak^\alpha+o(k^\alpha),\quad a\neq 0,\qquad(1.3)$$

for some $\alpha>\sigma-1$; this implies that Levinson's theorem takes the form

$$\delta_l(0)-\delta_l(\infty)=\pi(n_l+\alpha/2),\qquad(1.4)$$

where $\delta_l(k)$ denotes the lth phase shift and n_l is the number of negative eigenvalues for angular momentum l. Thus, if $\alpha\neq 1$ $(l=0)$ or $\alpha\neq 2$ $(l\geqslant 1)$, we would get a modified Levinson's theorem. However, we should be aware of the possibility that if we simply treat (1.3) as a special case of (1.2) we may miss some information that specifically pertains to the case when $F_l(0)=0$. Indeed, it is known that if $q\in L^1_1$ and $F_l(0)=0$, then we always have $\alpha=1$ when $l=0$ and $\alpha=2$ when $l=1$. A proof for $l=0$ (and a hint of how to proceed when $l\geqslant 1$) can be found in the work of Marchenko[2] (for $l=0$ a proof also follows from Ref. 3, Appendix I). Regarding (1.2) this leads to the question of whether the given error estimate is optimal for the class L^1_σ and of how the large-r behavior of $q(r)$ is reflected in the small-k behavior of $F_l(k)$.

The paper is organized as follows. In Sec. II we explain the notation, collect some preliminary material, and state Lemma (2.1), which is needed in the later sections. The proof is given in the Appendix.

Section III is devoted to the special case of inverse power-law potentials satisfying $q(r)\sim q_0r^{-2-\epsilon}$ as $r\to\infty$ with $0<\epsilon\leqslant 1$. For $l=0$ we obtain the leading behaviors of the remainder terms in (1.2) and (1.3) [see Theorem (3.1)]. This entails the leading behavior of the phase shift [see Corollary (3.2)] and extends, for $l=0$, previous results found by Keller and Levy,[4] who assumed $\epsilon>1$. There is a corresponding conjecture in the paper of Keller and Levy (Ref. 4, p. 59), but with the additional restriction that q be repulsive. We shall see that q can have arbitrary sign. Furthermore, we also consider the case when $F_0(0)=0$, which was not done in Ref. 4. It seems conceivable to us that results similar to those of Theorem (3.1) and Corollary (3.2) can be derived for arbitrary l (Keller and Levy also allowed $l>0$), but we have not checked the details. As a by-product of the analysis of power-law potentials we obtain precise information on the differentiability of the Jost solution $f_0(k,r)$ with respect to k at $k=0$ [see Corollary (3.3)]. This result clearly demonstrates why the Jost solutions cause problems in the analysis of $F_0(k)$ if $1\leqslant\sigma<2$, a fact that was also recognized in Ref. 1 (see Appendix C).

In Sec. IV we analyze the small-k behavior of $F_l(k)$ when $F_l(0)=0$ for arbitrary potentials with $1\leqslant\sigma<2$ $(l=0)$ or $1\leqslant\sigma<3$ $(l\geqslant 1)$ and we obtain Levinson's theorem for $\sigma=1$. Our proof makes essential use of Lemma (2.1), which allows us to bypass the differentiability problems associated with the Jost solutions. In fact, if Jost solutions are used (following the basic reference 5), then the stronger condition $q\in L^1_2$ seems to be unavoidable. This may have led to the wrong impression that this condition is actually necessary (Ref. 6, p. 23). Other recent proofs,[7,8] based on Sturmian arguments, also require that $r^2q(r)$ be integrable at infinity[8] or that $r^3q(r)\to 0$ as $r\to\infty$ [in Ref. 8 the special case $q(r)\sim q_0r^{-2}$ is also considered]. For a review of the subject see Bollé.[9] As already mentioned, the proofs in Refs. 2 and 3 work for $\sigma=1$, but they are based on the Marchenko equation (and on an inductive argument with respect to l when $l\geqslant 1$). Our proof is more direct in the sense that it is a refinement of Levinson's original proof,[10] which needed $\sigma=2$ at various places. Moreover, our method allows us to control the error terms.

0022-2488/88/010148-07$02.50

At the end of Sec. IV, we remark on how our results tie in with the threshold behavior of the eigenvalues[11,12] when they are born from the continuous spectrum as the coupling constant is increased.

Finally, we mention that the methods and results of this paper have extensions to the Dirac equation[13] and the Schrödinger equation on the line.[14] In the latter case we can, for example, prove continuity of the S matrix at $k=0$ for arbitrary potentials satisfying a L_1^1 condition on the line.

II. PRELIMINARIES

We consider the Schrödinger equation

$$-y'' + [l(l+1)/r^2]y + q(r)y = k^2 y \quad (l=0,1,...). \tag{2.1}$$

We always assume that $k>0$, except in Sec. IV, where we need $\operatorname{Im} k \geqslant 0$ in connection with Levinson's theorem. Let $y_l(k,r)$ denote the solution of (2.1) satisfying the boundary condition

$$y_l(k,r) \sim r^{l+1}, \quad \text{as } r\to 0. \tag{2.2}$$

Then y_l solves the integral equation

$$y_l(k,r) = y_{l0}(k,r) - \int_0^r dt\, g_l(k,r,t) q(t) y_l(k,t), \tag{2.3}$$

where

$$y_{l0}(r) = \Gamma(l+\tfrac{3}{2})(k/2)^{-l-1/2} r^{1/2} J_{l+1/2}(kr), \tag{2.4}$$

$$g_l(k,r,t) = \tfrac{1}{2}\pi(rt)^{1/2}(J_{l+1/2}(kr)Y_{l+1/2}(kt) - J_{l+1/2}(kt)Y_{l+1/2}(kr)). \tag{2.5}$$

Here, $J_{l+1/2}$ and $Y_{l+1/2}$ are the usual Bessel and Neumann functions and y_{l0} satisfies Eq. (2.1) with $q=0$. The Jost function $F_l(k)$ is defined by

$$F_l(k) = 1 + i\pi 2^{-l-3/2} k^{l+1/2}\left(\Gamma\left(l+\frac{3}{2}\right)\right)^{-1} \times \int_0^\infty dr\, r^{1/2} q(r) H_{l+1/2}^{(1)}(kr) y_l(k,r), \tag{2.6}$$

where $H_{l+1/2}^{(1)} = J_{l+1/2} + iY_{l+1/2}$ denotes the Hankel functions.

Moreover,

$$\delta_l(k) = -\arg F_l(k). \tag{2.7}$$

The zero-energy solution $y_l(r) \equiv y_l(0,r)$ will play an important role in this paper. Its main properties are the following: $y_l(r)$ is bounded at infinity if and only if $F_l(0) = 0$ and in that case it obeys

$$y_l(r) \sim A_l r^{-l}, \quad r\to\infty, \tag{2.8}$$

where

$$A_l = -\frac{1}{2l+1}\int_0^\infty dr\, r^{l+1} q(r) y_l(r) \neq 0. \tag{2.9}$$

Combining (2.2) and (2.8) we see that

$$|y_l(r)| \leqslant Cr^{l+1}(1+r)^{-2l-1}. \tag{2.10}$$

Here and subsequently C will denote various constants, although not necessarily the same at each appearance. If $F_l(0) \neq 0$, then $y_l(r)$ is unbounded and

$$y_l(r) \sim D_l r^{l+1}, \quad r\to\infty, \tag{2.11}$$

where

$$D_l = 1 + \frac{1}{2l+1}\int_0^\infty dr\, r^{-l} q(r) y_l(r). \tag{2.12}$$

Moreover,

$$D_l = F_l(0). \tag{2.13}$$

Also, notice that y_l is square integrable precisely when $F_l(0) = 0$ and $l \neq 0$. The above properties follow easily from the integral equation (2.3) and from (2.6). See, also, Ref. 1.

For our later proofs we need bounds on the difference $y_l(k,r) - y_l(r)$.

Lemma (2.1): Suppose that $q \in L_1^1$.

(i) If $l=0$, $F_0(0) \neq 0$, then

$$|y_0(k,r) - y_0(r)| \leqslant C_\delta r[kr/(1+kr)]^\delta \tag{2.14}$$

and if $F_0(0) = 0$, then

$$|y_0(k,r) - y_0(r)| \leqslant C_\delta [kr/(1+kr)]^\delta, \tag{2.15}$$

where $0 \leqslant \delta \leqslant 2$.

(ii) If $l \geqslant 1$, $F_l(0) = 0$, then

$$|y_l(k,r) - y_l(r)| \leqslant Ck^2[r/1+kr)]^{l+1}. \tag{2.16}$$

For a proof, see the Appendix. Notice the absence of the factor r in (2.15) as compared to (2.14). In Sec. IV we need to use a second, linearly independent solution $\tilde{y}_l(r)$ of (2.1) for $k=0$ if $l \geqslant 1$, $F_l(0) = 0$. We choose $\tilde{y}_l$ such that $y_l\tilde{y}_l' - y_l'\tilde{y}_l = 1$. Then

$$\tilde{y}_l(r) \sim \tilde{A}_l r^{l+1}, \quad r\to\infty, \tag{2.17}$$

where

$$(2l+1)A_l\tilde{A}_l = 1 \tag{2.18}$$

and

$$\tilde{y}_l(r) \sim -[1(2l+1)]r^{-l}, \quad r\to 0. \tag{2.19}$$

Hence

$$|\tilde{y}_l(r)| \leqslant Cr^{-l}(1+r)^{2l+1}. \tag{2.20}$$

Equations (2.17) and (2.19) follow easily from the representation

$$\tilde{y}_l(r) = y_l(r)\int_{r_0}^r dr\, y_l^{-2}(r) + \rho_l y_l(r),$$

where r_0 is at our disposal and ρ_l is a suitable constant (depending on r_0). The asymptotic relations (2.2), (2.8), (2.11), (2.17), and (2.19) may all be differentiated.

III. THE CASE $q(r) \sim q_0 r^{-2-\epsilon}$, $0<\epsilon\leqslant 1$ ($l=0$)

Theorem (3.1): Suppose that, as $r\to\infty$, $q(r) \sim q_0 r^{-2-\epsilon}$, $0<\epsilon\leqslant 1$.

(i) If $0<\epsilon<1$, $F_0(0)\neq 0$, then, as $k\to 0$,

$$F_0(k) = F_0(0) + a_0 e^{-(i/2)\pi\epsilon} F_0(0) k^\epsilon + o(k^\epsilon), \tag{3.1}$$

where

$$a_0 = -q_0 2^\epsilon(\epsilon(\epsilon+1))^{-1}\Gamma(1-\epsilon). \tag{3.2}$$

(ii) If $0<\epsilon<1$, $F_0(0) = 0$, then, as $k\to 0$,

$$F_0(k) = -ikA_0 + iA_0 a_0 e^{-(i/2)\pi\epsilon} 2^{-\epsilon} k^{\epsilon+1} + o(k^{\epsilon+1}). \tag{3.3}$$

(iii) If $\epsilon = 1$, then, as $k \to 0$,

$$F_0(k) = \begin{cases} F_0(0) - iq_0F_0(0)k\ln k + o(k\ln k), & F_0(0) \neq 0, \qquad (3.4a) \\ -ikA_0 + (A_0/2)q_0k^2\ln k + o(k^2\ln k), & F_0(0) = 0. \qquad (3.4b) \end{cases}$$

Relation (3.1) shows that in any class L_σ, $1 \leqslant \sigma < 2$, we can find a q such that $F_0(k) - F_0(0)$ vanishes like $k^{\sigma-1+\delta}$, with $\delta > 0$ as small as we wish. In this sense, when $F_0(0) \neq 0$, the remainder estimate in (1.2) is optimal. Theorem (3.1) has the following implications about the phase shifts.

Corollary (3.2): Under the assumptions of Theorem (3.1), if $0 < \epsilon < 1$, then

$$\delta_0(k) = \begin{cases} a_0k^\epsilon\sin(\pi\epsilon/2) + o(k^\epsilon), & F_0(0) \neq 0, \qquad (3.5a) \\ \pi/2 - a_02^{-\epsilon}k^\epsilon\sin(\pi\epsilon/2) + o(k^\epsilon), & F_0(0) = 0, \qquad (3.5b) \end{cases}$$

and, if $\epsilon = 1$, then

$$\delta_0(k) = \begin{cases} q_0k\ln k + o(k\ln k), & F_0(0) \neq 0, \qquad (3.6a) \\ \pi/2 - (q_0/2)k\ln k + o(k\ln k), & F_0(0) = 0. \qquad (3.6b) \end{cases}$$

These relations are all understood to hold mod(π). Corollary (3.2) follows from Theorem (3.1) and (2.7).

Proof of Theorem (3.1): (i) $F_0(0) \neq 0$, $0 < \epsilon < 1$. We break the integral in (2.6) into three parts:

$$F_0(k) = F_0(0) + \int_0^\infty dr(e^{ikr} - 1)q(r)y_0(r) + \int_0^\infty dr\, e^{ikr}q(r)(y_0(k,r) - y_0(r)), \qquad (3.7)$$

where we have also used (2.12) and (2.13). We denote the first integral on the rhs by I_1 and the second by I_2. Considering I_1, it is easy to show that the leading behavior of I_1 as $k \to 0$ is determined completely by the asymptotic forms for q and y_0 as $r \to \infty$. On substituting D_0r for y_0 and $q_0r^{-2-\epsilon}$ for $q(r)$ and changing variables, $u = kr$, we obtain

$$I_1 = k^\epsilon D_0q_0\int_0^\infty du(e^{iu} - 1)u^{-1-\epsilon} + o(k^\epsilon). \qquad (3.8)$$

Next we consider I_2. For $l = 0$, (2.3) reads as

$$y_0(k,r) = \frac{\sin kr}{k} + \frac{1}{k}\int_0^r dt\sin k(r-t)q(t)y_0(k,t) \qquad (3.9)$$

and

$$y_0(r) = r + \int_0^r dt(r-t)q(t)y_0(t). \qquad (3.10)$$

From this we deduce that

$$\begin{aligned} y_0(k,r) - y_0(r) &= r\left(\frac{\sin kr}{kr} - 1\right)\left(1 + \int_0^r dt\, q(t)y_0(t)\right) \\ &\quad + \frac{\sin kr}{k}\int_0^r dt\,(\cos kt - 1)q(t)y_0(t) \\ &\quad - \frac{1}{k}(\cos kr - 1)\int_0^r dt\sin ktq(t)y_0(t) \\ &\quad - \frac{1}{k}\int_0^r dt(\sin kt - kt)q(t)y_0(t) \\ &\quad + \frac{1}{k}\int_0^r dt\, q(t)\sin k(r-t)(y_0(k,t) - y_0(t)). \end{aligned} \qquad (3.11)$$

We denote the five terms on the rhs by $A_1(k,r),\ldots,A_5(k,r)$, respectively. Upon inserting $A_1(k,r)$ into the expression for I_2, we see that

$$\int_0^\infty dr\, e^{ikr}q(r)A_1(k,r) = k^\epsilon D_0q_0\int_0^\propto du\, e^{iu}\left(\frac{\sin u}{u} - 1\right)u^{-\epsilon-1} + o(k^\epsilon). \qquad (3.12)$$

Next we estimate A_2, A_3, and A_4. Let $\beta\in(\epsilon,2\epsilon)$. By using elementary estimates we deduce that

$$|A_j(k,r)| \leqslant Ck^\beta r^{1+\beta/2}\int_0^r dt\,|q(t)|t^{\beta/2}|y_0(t)| \leqslant Ck^\beta r^{1+\beta/2}, \quad j = 2,3,4, \qquad (3.13)$$

on account of the linear growth of y_0. This, in turn, implies

$$\left|\int_0^\infty dr\, e^{ikr}q(r)A_j(k,r)\right| \leqslant Ck^\beta, \quad j = 2,3,4. \qquad (3.14)$$

The term $A_5(k,r)$ is estimated by using (2.14) with $\delta = \beta$, so that

$$|A_5(k,r)| \leqslant Ck^\beta r^{1+\beta/2}\int_0^r dt\,|q(t)|t^{1+\beta/2}$$

and thus

$$\left|\int_0^\infty dr\, e^{ikr}q(r)A_5(k,r)\right| \leqslant Ck^\beta. \qquad (3.15)$$

Thus the contributions from A_2 through A_5 to I_2 are $o(k^\epsilon)$. Adding (3.8) and (3.12) and computing the remaining integral yields the second term on the rhs of (3.1).

(ii) $F_0(0) = 0$, $0 < \epsilon < 1$. We again use (3.7). Since now y_0 is bounded at infinity, we obtain, using (2.8) and (2.9),

$$I_1 = -ikA_0 + k^{\epsilon+1}A_0q_0 \cdot \int_0^\infty du(e^{iu} - 1 - iu)u^{-\epsilon-2} + o(k^{\epsilon+1}). \qquad (3.16)$$

In this case, however, $A_1(k,r)$ does not contribute to the $k^{\epsilon+1}$ term. In fact, since $D_0 = 0$, we can write

$$A_1(k,r) = -r\left(\frac{\sin kr}{kr} - 1\right)\int_r^\infty dt\, q(t)y_0(t) \qquad (3.17)$$

and thus

$$|A_1(k,r)|\leqslant Ck^{\gamma}r^{1/2+\gamma/2}\int_r^{\infty} dt\,|q(t)|t^{1/2+\gamma/2}|y_0(t)|, \tag{3.18}$$

where $1+\epsilon<\gamma<\min(2,1+2\epsilon)$, so that the contribution of A_1 to I_2 is $O(k^{\gamma})$. Also, we have

$$|A_j(k,r)|\leqslant Ck^{\gamma}r^{1/2+\gamma/2}\int_0^{r} dt\,|q(t)|t^{1/2+\gamma/2}|y_0(t)|,$$

$$j=2,3,4, \tag{3.19}$$

and a similar estimate for A_5 in view of inequality (2.15) with $\delta=\gamma$. Thus the contributions from A_1 through A_5 to I_2 are $O(k^{\gamma})=o(k^{\epsilon+1})$. Evaluating the integral in (3.16) yields (3.3).

(iii) If $\epsilon=1$, $F_0(0)\neq 0$, then

$$I_1=-iq_0D_0k\ln k+o(k\ln k). \tag{3.20}$$

Moreover,

$$\int_0^{\infty} dr\,e^{ikr}q(r)A_1(k,r)$$

$$\sim D_0q_0k\int_0^{\infty} du\,e^{iu}\frac{\sin u-u}{u^3}=O(k). \tag{3.21}$$

The contributions from A_2 through A_5 are $O(k^{\beta})$, $\beta\in(1,2)$. Remembering (2.13), we arrive at (3.4a). If $F_0(0)=0$, then

$$I_1=-ikA_0+(A_0/2)q_0k^2\ln k+o(k^2\ln k). \tag{3.22}$$

In the estimates for A_j ($j=1,\ldots,5$) we may, of course, choose $\gamma=1+\epsilon=2$ [see (3.18)] and we see that $I_2=O(k^2)$, whence (3.4b). Theorem (3.1) is proved.

Next we turn to the differentiability properties of the Jost solution $f_0(k,r)$ at $k=0$, where $f_0(k,r)$ denotes the solution of (2.1) defined by the boundary condition

$$\lim_{r\to\infty} e^{-ikr}f_0(k,r)=1. \tag{3.23}$$

Corollary (3.3): Assume that $q(r)\sim q_0r^{-2-\epsilon}$ as $r\to\infty$, $0<\epsilon\leqslant 1$ ($q_0\neq 0$). Then $f_0(k,r)$ is differentiable at $k=0$ if and only if $f_0(0,r)=0$.

Proof: For any $r=r_0>0$ we have

$$f_0(k,r_0)=e^{ikr_0}+\int_{r_0}^{\infty} dt\,e^{ikt}q(t)y(k,t;r_0), \tag{3.24}$$

where $y(k,r;r_0)$ solves (2.1) for $r>r_0$ with $y(k,r_0;r_0)=0$, $y'(k,r_0;r_0)=1$. The integral (3.24) can be analyzed in the same way as the integral (3.7) and we obtain the analog of Theorem (3.1) with respect to the interval $r>r_0$. In other words, $f_0(k,r_0)$ is differentiable at $k=0$ if and only if $f_0(0,r_0)=0$, which is the assertion of Corollary (3.3). Under the stronger assumption that $q\in L^1_2$, we know that $f_0(k,r)$ is continuously differentiable with respect to k at $k=0$ for any r.[1]

IV. THE CASE $F_l(0)=0$, LEVINSON'S THEOREM, AND THRESHOLD BEHAVIOR

Here we prove the following theorem.

Theorem (4.1): Suppose that $q\in L^1_{\sigma}$ and that (i) $l=0$, $F_0(0)=0$, $1\leqslant\sigma<2$, then, as $k\to 0$,

$$F_0(k)=-iA_0k+o(k^{\sigma}); \tag{4.1}$$

or (ii) $l\geqslant 1$, $F_l(0)=0$, $1\leqslant\sigma<3$, then, as $k\to 0$,

$$F_l(k)=c_lk^2+\begin{cases}o(k^{\sigma+1}), & 1\leqslant\sigma<3,\quad l\geqslant 2 \text{ or } 1\leqslant\sigma<2,\quad l=1,\\ O(k^3), & 2\leqslant\sigma<3,\ l=1,\end{cases} \tag{4.2}$$

where

$$c_l=-\frac{\|y_l\|^2}{(2l+1)A_l}=-\frac{(\partial F_l/\partial\lambda)(0;1)\|y_l\|^2}{(qy_l,y_l)}. \tag{4.3}$$

Here and subsequently (,) denotes the L^2 inner product and $\|\ \|$ denotes the L^2 norm. In (4.3) $F_l(k;\lambda)$ is the Jost function for (2.1) with q replaced by λq ($\lambda\in\mathbb{R}$). The second equation in (4.3) allows us to establish the connection with the threshold coupling constant behavior of the eigenvalues (see below).

Proof: (i) For $l=0$ we need only look back at the proof of Theorem (3.1), Eq. (3.7). We have

$$I_1=iA_0k+\int_0^{\infty} dr(e^{ikr}-1-ikr)q(r)y_0(r). \tag{4.4}$$

The integrand is $O(k^2)$ and dominated by $Ck^{\sigma}r^{\sigma}|q(r)|\,|y_0(r)|$; hence by dominated convergence the integral is $o(k^{\sigma})$. Thus

$$I_1=-iA_0k+o(k^{\sigma}). \tag{4.5}$$

From (2.15) with $\delta=\sigma$,

$$|I_2|\leqslant Ck^{\sigma}\int_0^{\infty} dr|q(r)|\left(\frac{r}{1+kr}\right)^{\sigma},$$

so again by dominated convergence [since $y_0(k,r)-y_0(r)=O(k^2)$ for fixed r],

$$I_2=o(k^{\sigma}). \tag{4.6}$$

This establishes (4.1).

(ii) The case when $l\geqslant 1$ is complicated by the fact that part of the leading contribution comes from the analog of I_2. We begin by splitting $F_l(k)$ as

$$\begin{aligned}F_l(k)={}&\alpha_lk^{l+1/2}\int_0^{\infty} dr\,r^{1/2}q(r)(H^{(1)}_{l+1/2}(kr)\\&-\widetilde{H}^{(1)}_{l+1/2}(kr))\,y_l(r)\\&+\alpha_lk^{l+1/2}\int_0^{\infty} dr\,r^{1/2}q(r)(H^{(1)}_{l+1/2}(kr)\\&-\widetilde{H}^{(1)}_{l+1/2}(kr))(y_l(k,r)-y_l(r))\\&+\alpha_lk^{l+1/2}\int_0^{\infty} dr\,r^{1/2}q(r)\\&\times\widetilde{H}^{(1)}_{l+1/2}(kr)(y_l(k,r)-y_l(r))\end{aligned} \tag{4.7}$$

and we denote the three terms on the rhs by B_1, B_2, and B_3, respectively. Here

$$\alpha_l=i\pi 2^{-l-3/2}(\Gamma(l+\tfrac{3}{2}))^{-1} \tag{4.8}$$

and $\widetilde{H}^{(1)}_{l+1/2}(kr)$ is the leading term of $H^{(1)}_{l+1/2}(kr)$ as $r\to 0$, i.e.,

$$\widetilde{H}^{(1)}_{l+1/2}(kr)=\beta_l(kr)^{-l-1/2}, \tag{4.9}$$

where

$$\beta_l = 1/(2l+1)\alpha_l. \tag{4.10}$$

Then

$$|H^{(1)}_{l+1/2}(kr) - \widetilde{H}^{(1)}_{l+1/2}(kr)| \leqslant c(kr)^{-l+3/2}(1+kr)^{l-2}. \tag{4.11}$$

We consider B_2 first. Upon inserting (4.11) and (2.16) into B_2 we see that

$$\begin{aligned}|B_2| &\leqslant Ck^4 \int_0^\infty dr \frac{r^3|q(r)|}{(1+kr)^3} \\ &\leqslant Ck^{\sigma+1} \int_0^\infty dr\, r^\sigma |q(r)| \frac{(kr)^{3-\sigma}}{(1+kr)^3} \\ &= o(k^{\sigma+1}).\end{aligned} \tag{4.12}$$

To analyze B_1, we expand $H^{(1)}_{l+1/2}$ one term further:

$$H^{(1)}_{l+1/2}(kr) = \beta_l(kr)^{-l-1/2} + \gamma_l(kr)^{-l+3/2} + R_l(kr), \tag{4.13}$$

where

$$\gamma_l = \beta_l/2(2l-1). \tag{4.14}$$

Now, $R_l(k,r)$ obeys the following estimates. For $l=1$,

$$|R_1(k,r)| \leqslant C(kr)^{3/2}(1+kr)^{-1} \tag{4.15}$$

and, for $l \geqslant 2$,

$$|R_l(k,r)| \leqslant C(kr)^{7/2-l}(1+kr)^{l-4}. \tag{4.16}$$

By using (2.10) it is easy to see that the contribution from $R_l(k,r)$ to B_1 is $O(k^3)$ if $l=1$ and $O(k^4)$ if $l \geqslant 2$, provided only $\sigma = 1$. Splitting off the leading term, we obtain

$$B_1 = \alpha_l \gamma_l k^2 \int_0^\infty dr\, r^{2-l} q(r) y_l(r) + \begin{cases} O(k^3), & l=1, \\ O(k^4), & l \geqslant 2. \end{cases} \tag{4.17}$$

It remains for us to consider B_3. We make use of another representation for $y_l(k,r)$, which we obtain by applying the variation of parameter formula to (2.1), namely

$$y_l(k,r) = y_l(r) + k^2 u_l(r) + T_l(k,r), \tag{4.18}$$

where

$$u_l(r) = \int_0^r dt\, h_l(r,t) y_l(t), \tag{4.19}$$

$$h_l(r,t) = y_l(r)\tilde{y}_l(t) - \tilde{y}_l(r) y_l(t), \tag{4.20}$$

and

$$T_l(k,r) = k^2 \int_0^r dt\, h_l(r,t)(y_l(k,t) - y_l(t)). \tag{4.21}$$

Here $\tilde{y}_l$ is the solution discussed in Sec. II, (2.17)–(2.20). From the properties of y_l and $\tilde{y}_l$ we infer that

$$|h_l(r,t)| \leqslant Cr^{l+1} t^{-l}, \quad t \leqslant r, \tag{4.22}$$

$$u_l(r) \sim -\widetilde{A}_l \|y_l\|^2 r^{l+1}, \quad r \to \infty, \tag{4.23}$$

$$u_l(r) \sim \text{const}\, r^{l+3}, \quad r \to 0. \tag{4.24}$$

Moreover, by (4.22) and (2.16) we obtain the bound

$$|T_l(k,r)| \leqslant Ck^4 r^{l+3}. \tag{4.25}$$

Now we put $R = 1/k$ and write

$$\begin{aligned}B_3 &= \alpha_l k^{l+1/2} \int_R^\infty dr\, r^{1/2} q(r) \widetilde{H}_{l+1/2}(kr) \\ &\quad \times (y_l(k,r) - y_l(r)) + \alpha_l k^{l+1/2} \int_0^R dr\, r^{1/2} q(r) \\ &\quad \times \widetilde{H}_{l+1/2}(kr)(y_l(k,r) - y_l(r)) = J_1 + J_2.\end{aligned} \tag{4.26}$$

The term J_1 is estimated by means of (2.16) and (4.9):

$$\begin{aligned}|J_1| &\leqslant Ck^2 \int_R^\infty dr\, r|q(r)| \\ &\leqslant Ck^{1+\sigma} \int_R^\infty dr\, r^\sigma |q(r)| = o(k^{\sigma+1}).\end{aligned} \tag{4.27}$$

Now we write J_2 [using (4.9)] as

$$\begin{aligned}J_2 &= \alpha_l \beta_l k^2 \int_0^R dr\, q(r) r^{-l} u_l(r) \\ &\quad + \alpha_l \beta_l \int_0^R dr\, q(r) r^{-l} T_l(k,r).\end{aligned} \tag{4.28}$$

We split the second integral in Eq. (4.28) into two, with one going from 0 to $R^{1/2}$ and the other from $R^{1/2}$ to R and estimate them by using (4.25):

$$\begin{aligned}&\left| \int_0^R dr\, q(r) r^{-l} T_l(k,r) \right| \\ &\leqslant Ck^{(5+\sigma)/2} \int_0^{\sqrt{R}} dr |q(r)| r^\sigma \\ &\quad + Ck^{\sigma+1} \int_{\sqrt{R}}^R dr |q(r)| r^\sigma = o(k^{\sigma+1}).\end{aligned} \tag{4.29}$$

We write the first term on the rhs of (4.28) as

$$\begin{aligned}&\alpha_l \beta_l k^2 \int_0^\infty dr\, q(r) r^{-l} u_l(r) \\ &\quad - \alpha_l \beta_l k^2 \int_R^\infty dr\, q(r) r^{-l} u_l(r)\end{aligned} \tag{4.30}$$

and observe that here the second term is bounded by [use (4.23)]

$$Ck^{\sigma+1} \int_R^\infty dr |q(r)| r^\sigma = o(k^{\sigma+1}). \tag{4.31}$$

Thus

$$B_3 = \alpha_l \beta_l k^2 \int_0^\infty dr\, q(r) r^{-l} u_l(r) + o(k^{\sigma+1}). \tag{4.32}$$

Thus from (4.12), (4.17), and (4.32) we obtain (4.2) with

$$\begin{aligned}c_l &= \alpha_l \gamma_l \int_0^\infty dr\, r^{2-l} q(r) y_l(r) \\ &\quad + \alpha_l \beta_l \int_0^\infty dr\, r^{-l} q(r) u_l(r).\end{aligned} \tag{4.33}$$

We must still transform c_l into (4.3). To this end, we observe that

$$-u_l'' + qu_l + [l(l+1)/r^2] u_l = y_l. \tag{4.34}$$

Upon multiplying Eq. (4.34) by r^{-l} and integrating by parts twice, we obtain the relation

$$\int_0^\infty dr\, r^{-l} y_l(r) = (2l+1)\widetilde{A}_l \|y_l\|^2 + \int_0^\infty dr\, q(r) r^{-l} u_l(r). \qquad (4.35)$$

The first term on the rhs comes from $r = \infty$ because of (4.23) and (2.8). In a similar manner we deduce from Eq. (2.1) with $k = 0$ that

$$2(1-2l)\int_0^\infty dr\, y_l(r) r^{-l} = \int_0^\infty dr\, q(r) y_l(r) r^{2-l}. \qquad (4.36)$$

By using (4.36), (4.35), (4.8), (4.10), (4.14), and (2.18) we obtain

$$c_l = -\widetilde{A}_l \|y_l\|^2 = -\|y_l\|^2/(2l+1)A_l, \qquad (4.37)$$

which is the first relation in (4.3). To establish the second relation in (4.3) we proceed as follows. Let G_l denote the integral operator having kernel $g_l(0,r,r')$. Then (2.3) for $k = 0$ becomes

$$y_l = r^{l+1} - G_l q y_l. \qquad (4.38)$$

Since $F_l(0) = 0$, i.e., $D_l = 0$, we also have

$$y_l = A_l r^{-l} - G_l^* q y_l, \qquad (4.39)$$

where G_l^* is the adjoint of G_l. Now we introduce a coupling constant λ [i.e., we replace q by λq in (2.1)] and denote the corresponding zero-energy solution and Jost function by $y_l(r;\lambda)$ and $F_l(k;\lambda)$, respectively. We have that $y_l(r;1) = y_l(r)$ and $F_l(k;1) = F_l(k)$, $F_l(0;1) = 0$. We also put

$$y_{l;\lambda}(r) = \frac{\partial y_l}{\partial z}(r;z)\bigg|_{z=\lambda}, \qquad (4.40)$$

$$F_{l;\lambda}(k) = \frac{\partial F_l}{\partial z}(k;z)\bigg|_{z=\lambda}. \qquad (4.41)$$

Then by (2.6) and (2.12),

$$\begin{aligned} F_{l;1}(0) &= \frac{1}{2l+1}\int_0^\infty dr\, r^{-l} q(r) y_l(r) \\ &\quad + \frac{1}{2l+1}\int_0^\infty dr\, r^{-l} q(r) y_{l;1}(r) \\ &= -1 + \frac{1}{2+1}\int_0^\infty dr\, r^{-l} q(r) y_{l;1}(r), \qquad (4.42) \end{aligned}$$

$$y_{l;1} = -G_l q y_l - G_l q y_{l;1}. \qquad (4.43)$$

From Eq. (4.43) $y_{l;1}$ can be obtained by iteration. To simplify the notation in the following calculations, we freely use the notation (f,g) even if f and g are not in L^2, but fg is in L^1. Then, by (4.38), (4.39), (4.43), and (2.9),

$$\begin{aligned} A_l(qr^{-l}, y_{l;1}) &= (q(1+G_l^* q)y_l, y_{l;1}) \\ &= ((1+qG_l^*)qy_l, y_{l;1}) \\ &= (qy_l(1+G_l q)y_{l;1}) \\ &= -(qy_l, G_l q y_l) \\ &= (qy_l, y_l) - (qy_l, r^{l+1}) \\ &= (qy_l, y_l) + (2l+1)A_l. \qquad (4.44) \end{aligned}$$

Thus

$$F_{l;1}(0) = (qy_l, y_l)/(2l+1)A_l \qquad (4.45)$$

or

$$c_l = -F_{l;1}(0)\|y_l\|^2/(qy_l, y_l), \qquad (4.46)$$

which is the desired second form for c_l. The proof of Theorem (4.1) is complete.

Relation (4.45) also holds when $l = 0$ [recall that y_0 is bounded so that (qy_0, y_0) exists].

The proof of part (i) given here is a simpler version of a proof that appears in Ref. 15.

A. Levinson's theorem

Since if $F_l(0) = 0$, then $k^{-1}F_0(k)$ or $k^{-2}F_l(k)$ $(l \geqslant 1)$ tends to a finite limit as $k \downarrow 0$ and $F_l(-k) = \overline{F_l(k)}$; the same limits are approached as $k \uparrow 0$. Moreover, $F_l(k)$ is analytic for $\operatorname{Im} k > 0$ and continuous for $\operatorname{Im} k \geqslant 0$. Then by a Phragmen–Lindelöf theorem[16] (the required exponential bound is established easily) these same limits are approached as $k \to 0$ from the upper half-plane. Hence by the usual contour argument, the contribution from the point $k = 0$ leads to $\alpha = 1\,(l = 0)$ or $\alpha = 2\,(l \geqslant 1)$ in (1.4) [when $F_l(0) = 0$] and the ordinary Levinson theorem holds.

B. Threshold behavior

Suppose that $q \in L_1^1$ and $F_l(0;1) = 0$. Upon expanding $F_l(k;\lambda)$ near $k = 0$ and $\lambda = 1$ (F_l is analytic in λ) and using (4.1), (4.2), and (4.46) we conclude that there is a function $k(\lambda)$ obeying

$$F_l(k(\lambda), \lambda) = 0 \qquad (4.47)$$

and

$$k(\lambda) = -iA_0^{-2}(qy_0, y_0)(\lambda - 1) + o(\lambda - 1), \quad l = 0, \qquad (4.48)$$

$$k(\lambda) = i[\,|(qy_l, y_l)|^{1/2}/\|y_l\|\,](\lambda - 1)^{1/2} + o((\lambda - 1)^{1/2}), \quad l \geqslant 1. \qquad (4.49)$$

Since $(qy_l, y_l) < 0$ (4.47) means that $k^2(\lambda)$ is a negative eigenvalue of Eq. (2.1), converging to 0 as $\lambda \downarrow 1$. The manner in which it does so is, for $l \geqslant 1$, in agreement with a general theorem found by Simon[12] and, for $l = 0$, consistent with related results[11] but at least $\sigma = 2$ was required in Ref. 11.

Note added in proof: We were unaware of the book by V. V. Babikov [*The Variable Phase Method in Quantum Mechanics* (Nauka, Moscow, 1968) (in Russian)]. It contains some results about the low-energy behavior of the phase shift for inverse power-law potentials (p. 121). We wish to thank D. Bollé and F. Gesztesy for pointing this reference out to us.

ACKNOWLEDGMENTS

We wish to thank P. Deift, F. Gesztesy, and R. Newton for useful discussions.

APPENDIX: PROOF OF LEMMA (2.1)

(i) Clearly, it suffices to prove (2.14) and (2.15) when $\delta = 2$. Since a proof of (2.14) is given in Ref. 1 we omit it here. Thus we turn to the case $F_0(0) = 0$. Here we use the decomposition (3.11) along with (2.12) and (2.13) (i.e., $D_0 = 0$) and rewrite the term $A_1(k,r)$ as

$$A_1(k,r) = -r\left(\frac{\sin kr}{kr} - 1\right)\int_r^\infty dt\, q(t)y_0(t). \tag{A1}$$

Thus

$$|A_1(k,r)| \leqslant C\,[kr/(1+kr)]^2. \tag{A2}$$

Similarly, we easily see by means of elementary estimates such as $|\sin z - z| \leqslant cz^3/(1+z)^2$ and by using the monotonicity of $z/1+z$ that

$$|A_j(k,r)| \leqslant C\,[kr/(1+kr)]^2, \quad j = 2,3,4. \tag{A3}$$

Also,

$$|A_5(k,r)| \leqslant \frac{Cr}{1+kr}\int_0^r dt\,|q(t)|\,|y_0(k,t) - y_0(t)|. \tag{A4}$$

Thus, letting $u(k,r) = y_0(k,r) - y_0(t)$, we have

$$|u(k,r)| \leqslant c\left(\frac{kr}{1+kr}\right)^2 + c\frac{r}{1+kr}\int_0^r dt\,|q(t)|\,|u(k,t)|. \tag{A5}$$

By applying Gronwall's lemma, we obtain

$$|u(k,r)| \leqslant C\,[kr/(1+kr)]^2, \tag{A6}$$

whence (2.15).

(ii) The proof is similar in spirit to case (i). By using $D_l = 0$, we may write

$$y_l(k,r) - y_l(r) = I_1 + \cdots + I_5, \tag{A7}$$

where

$$I_1 = -2^{l-1/2}\Gamma(l+1/2)k^{-l-1/2}r^{1/2}(J_{l+1/2}(kr)$$
$$-\widetilde{J}_{l+1/2}(kr))\int_r^\infty dt\, t^{-l}q(t)y_l(t), \tag{A8}$$

$$I_2 = -\frac{\pi}{2}\int_0^r dt(rt)^{1/2}J_{l+1/2}(kr)(Y_{l+1/2}(kt)$$
$$-\widetilde{Y}_{l+1/2}(kt))q(t)y_l(t), \tag{A9}$$

$$I_3 = \frac{\pi}{2}\int_0^r dt(rt)^{1/2}(Y_{l+1/2}(kr)$$
$$-\widetilde{Y}_{l+1/2}(kr))\,J_{l+1/2}(kt)q(t)y_l(t), \tag{A10}$$

$$I_4 = \frac{\pi}{2}\int_0^r dt(rt)^{1/2}\widetilde{Y}_{l+1/2}(kr)(J_{l+1/2}(kt)$$
$$-\widetilde{J}_{l+1/2}(kt))q(t)y_l(t), \tag{A11}$$

$$I_5 = -\int_0^r dt\, g_l(k,r,t)q(t)(y_l(k,t) - y_l(t)), \tag{A12}$$

and where $\widetilde{J}_{l+1/2}$, $\widetilde{Y}_{l+1/2}$ denote the leading parts of $J_{l+1/2}$, $Y_{l+1/2}$ as $r \to 0$, respectively. Explicitly,

$$\widetilde{J}_{l+1/2}(kr) = (\Gamma(l+\tfrac{3}{2}))^{-1}(kr/2)^{l+1/2}, \tag{A13}$$

$$\widetilde{Y}_{l+1/2}(kr) = -2^{l+1/2}\pi^{-1}\Gamma(l+1/2)(kr)^{-l-1/2}. \tag{A14}$$

We note the estimates

$$|J_{l+1/2}(z)| \leqslant Cz^{l+1/2}(1+z)^{-l-1}, \quad z > 0, \tag{A15}$$

$$|J_{l+1/2}(z) - \widetilde{J}_{l+1/2}(z)| \leqslant Cz^{l+5/2}(1+z)^{-2}, \tag{A16}$$

$$|Y_{l+1/2}(z) - \widetilde{Y}_{l+1/2}(z)| \leqslant Cz^{-l+3/2}(1+z)^{l-2}, \tag{A17}$$

$$|g_l(k,r,t)| \leqslant Ck^{-1}[kr/(1+kr)]^{l+1}$$
$$\times[kt/(1+kt)]^{-l}, \quad t < r. \tag{A18}$$

By using these estimates and (2.10) we can check that

$$|I_i| \leqslant Ck^2[r/(1+kr)]^{l+1}, \quad i = 1,\ldots,4. \tag{A19}$$

Using (A18) to estimate I_5 and letting $u_l(k,r) = y_l(k,t) - y_l(t)$ we obtain

$$|u_l(k,r)| \leqslant Ck^2\left(\frac{r}{1+kr}\right)^{l+1} + Ck^{-1}\left(\frac{kr}{1+kr}\right)^{l+1}$$
$$\times\int_0^r dt\,|q(t)|\left(\frac{kt}{1+kt}\right)^{-l}|u_l(k,t)|. \tag{A20}$$

Hence by Gronwall's lemma,

$$|u_l(k,r)| \leqslant Ck^2[r/(1+kr)]^{l+1}. \tag{A21}$$

This proves Lemma (2.1).

[1] R. Newton, J. Math. Phys. **27**, 2720 (1986).

[2] V. A. Marchenko, *Sturm–Liouville Operators and Applications* (Birkhäuser, Basel, 1986).

[3] Z. S. Agranovich and V. A. Marchenko, *The Inverse Problem of Scattering Theory* (Gordon and Breach, New York, 1963).

[4] J. B. Keller and B. R. Levy, J. Math. Phys. **4**, 54 (1963).

[5] R. Newton, J. Math. Phys. **1**, 319 (1960).

[6] K. Chaden and P. C. Sabatier, *Inverse Problems in Quantum Scattering Theory* (Springer, New York, 1977).

[7] Z. R. Iwinski, L. Rosenberg, and L. Spruch, Phys. Rev. A **31**, 1229 (1985).

[8] Z. Q. Ma, J. Math. Phys. **26**, 1995 (1985).

[9] D. Bollé, in *Mathematics + Physics*, edited by L. Streit (World Scientific, Singapore, 1986).

[10] N. Levinson, Det. Kgl. Dan. Videnskabernes Selsk. Mat. Fys. Medd. **25** (10), 1 (1949).

[11] M. Klaus and B. Simon, Ann. Phys. (NY) **130**, 251 (1980).

[12] B. Simon, J. Func. Anal. **25**, 338 (1977).

[13] D. Hinton, M. Klaus, and K. Shaw, preprint, 1987.

[14] M. Klaus, preprint, 1987.

[15] M. Klaus, in *Proceedings of the Conference on Oscillation, Bifurcation and Chaos* (Can. Math Soc. Annual Seminar, Toronto, 1986).

[16] R. P. Boas, *Entire Functions* (Academic, New York, 1954).

Three interacting particles in one dimension: An algebraic approach

J. B. McGuire
Department of Physics, Florida Atlantic University, Boca Raton, Florida 33431

C. A. Hurst
Department of Mathematical Physics, University of Adelaide, Adelaide, SA 5001, Australia

(Received 19 May 1987; accepted for publication 2 September 1987)

An algebraic formulation of the problem of three particles in one dimension is given, where the particles interact with delta function potentials of arbitrary strength and have almost arbitrary mass. An algebraic formulation is taken to mean that the steps implied from formulation to solution involve finite algebra. The canonical example is equal mass particles interacting with equal strength delta function potentials, where the Bethe ansatz holds and the solution involves only sums of products of matrices with elements that are rational functions of a complex variable. When the Bethe ansatz fails the Sommerfeld diffraction ansatz is satisfied if a condition of internal consistency is met. This condition of internal consistency requires the solution to a Riemann–Hilbert functional equation with an algebraic coefficient. The solution to this functional equation is an analytic, but not generally a meromorphic function. It is demonstrated that an asymptotic solution may be constructed within the domain of algebraic functions.

I. INTRODUCTION

We shall consider here the quantum system of three particles of arbitrary mass in one dimension interacting with delta function potentials of arbitrary strength. Our goal is to show that these problems are exactly solvable. We shall take "exactly solvable" to mean that all of the algebra implied between formulation and solution is finite and therefore may, in principle, be evaluated. We take "exactly solved" to mean that this evaluation has been carried out. We shall leave the exact solution of various cases to subsequent work.

If the masses of the particles are equal and the delta function strengths are equal the problem is solvable.[1] The algebra of the problem is factorized by the Bethe ansatz. Gaudin[2] has extensively studied the problems solved by the Bethe ansatz technique. These problems are essentially all of the exactly solvable models of particle mechanics and statistical mechanics.

The Bethe ansatz technique is so pervasive in the exactly solvable models of mathematical physics that the algebraic consistency condition of the ansatz has been referred to as a condition of solvability, or a condition of complete integrability. The implication is that should the condition fail to be met, the problem is unsolvable. We shall demonstrate that for the class of problems under study here, this implication is incorrect.

Examination of the details of the factorization of the equal mass equal strength delta function problem helps to illuminate the true algebraic meaning of the Bethe ansatz consistency condition. Gaudin[2] provides a careful study of this point, which we briefly summarize here.

The Bethe ansatz of the equal mass, equal strength delta function problem assumes that the state function of the particles is given by a set of occupation numbers of plane wave states, and that these occupation numbers change according to two-particle amplitudes only when the coordinate separations of the pairs of particles reverse.

Consistency of this assumption is established by showing that the values of the occupation numbers for any permutation of the particles in state A (some permutation of the order of the particles along the line) conditioned upon given occupation numbers in state B is independent of the path taken from A to B. This consistency condition is often called a "star–triangle relation."

If a star–triangle relation is satisfied the problem is exactly solvable because the algebra required is finite. When the Bethe ansatz is satisfied the computation of the occupation numbers involves only a finite number of multiplications and additions of functions of parameters and dynamical variables. These algebraic operations, often called "transfer matrix methods," may, in principle, be carried out completely. Baxter[3] offers examples of problems solved by this technique.

In the case of three particles interacting with delta function potentials this consistency condition is satisfied only if the masses of the particles and the strengths of the delta function potentials are all equal. There are two results in the literature where the problems are exactly solved and the Bethe ansatz fails. The authors[4] analyzed an impenetrable case, where the three interacting particles were of arbitrary mass, but constrained to be in a fixed order along the one dimension. Gaudin and Derrida[5] analyzed a case where all masses are equal, two delta function strengths are equal, and one delta function is of zero strength.

In general outline this formulation will follow that of Ref. 4. In Sec. III we make an ansatz, the Sommerfeld[6] diffraction ansatz. It is an assumption of this ansatz that the analog of the occupation numbers are members of a certain class of analytic functions of the independent variables.

In Sec. IV we find that these occupation numbers satisfy a set of matrix difference equations. As in transfer matrix methods, a finite number of algebraic steps is required from the formulation to the solution of these difference equations. These algebraic steps are indicated in Sec. V. Many of the

0022-2488/88/010155-14$02.50

algebraic steps indicated cannot be done with pencil and paper by a human being. Computer assistance is required.

Generally, we will only indicate what algebra is to be done. Much of the mathematical justification for the indicated computations is left to later publication.

The Sommerfeld ansatz also requires that a condition of internal consistency be satisfied. This consistency condition requires that the analytic functions involved satisfy a Riemann–Hilbert functional equation. Functional equations of this type appear in Refs. 4 and 5, but the methods used in their solution are inadequate to deal with the algebraic structure of the general problem.

The methods developed in Sec. V are applicable to a class of problems where the problems of Refs. 4 and 5 appear as special cases. The sense in which these cases are special is algebraic. They are separated from the general case by being cases where the coefficient of the Riemann–Hilbert functional equation is an algebraic function whose Riemann surface is topologically genus zero or one. The genus of the algebraic coefficient of the general case can be any integer, and the whole character of the solution changes abruptly.

The solution to the matrix difference equations is most easily achieved in a particular basis. This basis is not the most convenient for the usual boundary conditions for particle scattering. Section VI deals with the form of integrals of the Sommerfeld ansatz that transform the solutions of the matrix difference equations into the amplitudes for physical processes. Section VII shows how to explicitly calculate those amplitudes for asymptotic boundary conditions.

II. FORMULATION OF THE PROBLEM

The Hamiltonian for three particles of masses m_1, m_2, m_3, interacting with delta function potentials of strength g_1, g_2, g_3, is

$$\mathrm{H} = \frac{p_1{}^2}{2m_1} + \frac{p_2{}^2}{2m_2} + \frac{p_3{}^2}{2m_3} + g_1\delta(x_1 - x_2) + g_2\delta(x_2 - x_3) + g_3\delta(x_1 - x_3).$$

Several papers[1,5] give the details of a transformation to the center of mass that reduces the problem to one with two independent variables. We will not repeat the transformation or its attendant algebra here, but will just remind the reader of the result.

The stationary state problem to be solved is

$$(\nabla^2 + k^2)\Psi = 0,$$

except upon lines in the two-dimensional state space where the delta function potentials act. These lines are shown in Fig. 1. The important features are as follows.

(1) *Each wedge corresponds to an ordering of the three particles along a line. The angle of opening of the wedge depends upon this ordering and upon the masses of the three particles.*

If m_l, m_c, m_r correspond, respectively, to the three masses in order left, center, right, the angle of opening of the corresponding wedge is

$$\alpha = \tan^{-1}[(m_l + m_c + m_r)m_c/m_l m_r]^{1/2}.$$

(2) *Delta function boundaries lie along radial lines where the wave function is nonanalytic.*

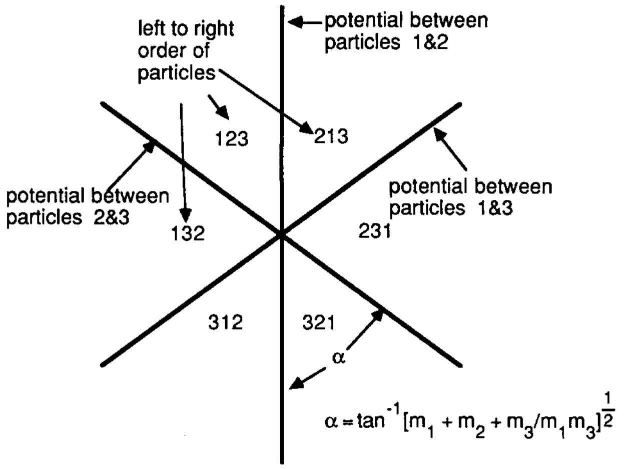

FIG. 1. The two-dimensional state space in the three-particle center of mass.

By integrating the differential equation along a line perpendicular to the boundary it is found that on this boundary the wave function satisfies a two-sided boundary condition. There is a discontinuity in the normal derivative that is equal to the value of the wave function times the strength of the delta function on the boundary, i.e.,

$$\left.\frac{\partial\Psi}{\partial\mathbf{n}}\right|_+ - \left.\frac{\partial\Psi}{\partial\mathbf{n}}\right|_- = g_k\Psi(0),$$

where g_k is the strength of the delta function along the k th boundary.

III. THE SOMMERFELD ANSATZ

A. The state function

We now make the Sommerfeld diffraction ansatz.[6] We assert that in each wedge between the delta function potentials the solution may be written as a path integral of the form

$$\Psi(r,\theta_k) = \int_C F_k(w,\theta_k)e^{ikr\cos w}\,dw,$$

where $F(w,\theta_k)$ is an analytic function of the two variables w,θ_k. The contour integral is computed over a path in the complex w plane which must be chosen to justify the following manipulations.

(1) We wish to replace differentiation with respect to r with differentiation with respect to w under the integral sign, and integrate by parts.

To justify this operation we require that the contour pass to infinity in two different places in the complex w plane and that the integral along this contour does not diverge. Under this restriction the partial differential equation in r, θ_k is satisfied if

$$\frac{\partial^2 F}{\partial w^2} - \frac{\partial^2 F}{\partial\theta_k^2} = 0,$$

which implies that

$$F = G_k(w + \theta_k) + H_k(w - \theta_k),$$

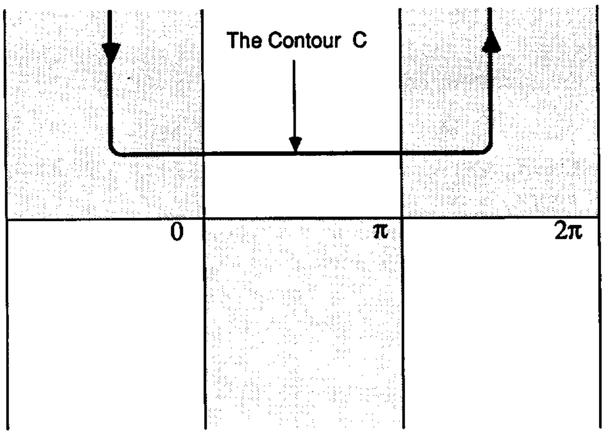

FIG. 2. The basic contour.

where $G_k(x)$, $H_k(x)$ are analytic functions of the single variable x. Here G_k and H_k are different functions in each region of the two-dimensional state space.

Under the same conditions it is possible to move the operation of differentiation with respect to the normal to a radial line to an operation under the integral. By the same sequence of operations (i.e., differentiate under the integral sign and integrate by parts) it may be shown that

$$\frac{\partial\Psi}{\partial\mathbf{n}} = \frac{\partial\Psi}{r\,\partial\theta_k}$$
$$= ik\int_C \sin w[G(w+\theta_k) - H(w-\theta_k)]e^{ikr\cos w}\,dw.$$

(2) We wish to be able to perform the integrals over G and H independently in either the calculation of Ψ or its normal derivative.

This requires that as θ_k varies through real values, no pole or branch of either G or H is crossed by the contour. We anticipate that all of the singularities of these functions will be at some finite distance from the real axis (i.e., none will be at infinity), and therefore we choose the contour so that it bypasses all of the singularities by always being greater than this distance from the real axis.

Figure 2 shows a contour in the complex w plane that satisfies both of these constraints. There are infinitely many other such contours, where $w \to w + 2n\pi$ in the upper half-plane, and $w \to 2l\pi - w$ in the lower half-plane.

B. Probability flux conservation

The Sommerfeld ansatz assumes that we may write an integral representation of a solution

$$\Psi(r,\theta) = \int_C G(w)e^{ikr\cos(w-\theta)}\,dw + \int_C H(w)e^{ikr\cos(w+\theta)}\,dw,$$

where Ψ, G, and H are all column vectors each of whose elements are associated with one of the regions of state space.

The probability flux through a circle whose center lies at the center of mass in state space is proportional to the flux matrix

$$\Phi = i\left[\Psi\times\left(\frac{\partial\Psi^*}{\partial r}\right) - \left(\frac{\partial\Psi}{\partial r}\right)\times\Psi^*\right],$$

where the $\times$ indicates the outer or tensor product of the column vectors.

The integral representation of this flux matrix contains several terms, each of which has a θ-dependent scalar kernel. A typical form for this kernel is

$$K(\theta) = \{\cos(w-\theta) + \cos(w'-\theta)\}$$
$$\times e^{ikr\{\cos(w-\theta) - \cos(w'-\theta)\}},$$

which may be rewritten as

$$K(\theta) = 2\cos(\tfrac{1}{2}(w+w') - \theta)$$
$$\times e^{-2ikr\sin(1/2(w-w'))\sin(1/2(w+w')-\theta)}.$$

Thus, for all cases, the probability flux matrix contains a scalar term of the form

$$K(\theta) = \frac{df}{d\theta}e^{iAf(0)},$$

where f is either $\sin\theta$ or $\cos\theta$. It is therefore true that

$$\int_0^{2\pi} K(\theta)d\theta = 0,$$

which provides a flux conservation theorem.

If a Sommerfeld integral representation exists, the flux into a full circle centered at the center of mass is zero. The probability to be within that circle is constant.

IV. FORMULATION OF THE DIFFERENCE EQUATIONS

A. The first difference equations

In order to present a single consistent formulation that is suitable for almost any masses of the particles and any strength of the delta function potentials, we shall study a generalization of the problem as presented.

Instead of the wedge structure of Fig. 1 with six wedges of varying angles, we consider a structure in which the circle is tesselated into N regions where the angle of opening is the same in each wedge, namely $2\pi/N$. We further assume that N is an even integer $N = 2m$, and therefore the angle of opening of is each wedge is π/m. Each pair of regions is separated by a delta function barrier, where the two-sided boundary condition must be satisfied. Each delta function boundary has associated with it a strength g_k, where $0<k<N-1$.

It is assumed that this tesselation represents no restriction, since any masses and strengths will fall arbitrarily close to this arrangement, provided that we make N sufficiently large and choose the appropriate delta function strengths to be zero.

The column vector Ψ_k is to be represented as

$$\Psi_k(r,\theta_k) = \int_C [G(w+\theta_k) + H(w-\theta_k)]e^{ikr\cos w}\,dw, \tag{4.1}$$

where each value of k is identified with one of the regions of the state space. The normal derivative is given by

$$\frac{\partial\Psi}{\partial\mathbf{n}} = \frac{\partial\Psi}{r\,\partial\theta_k}$$
$$= ik\int_C \sin w[G(w+\theta_k) - H(w-\theta_k)]e^{ikr\cos w}\,dw.$$

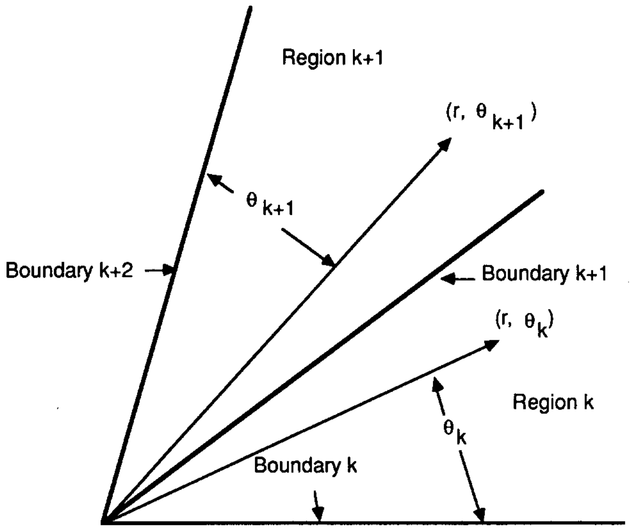

FIG. 3. Coordinate choices for two consecutive regions.

We will also choose the direction of increasing θ_k to alternate from wedge to wedge, as indicated in Fig. 3. This introduces an artificial but convenient distinction between even boundaries and odd boundaries. (See Fig. 3.)

The delta function boundary condition is applied successively at each boundary. The function and normal derivative integral representation of the Sommerfeld ansatz produce a set of equations that must be satisfied by the analytic functions G_k and H_k. We write a typical pair of equations assuming that boundary k separates region k from $k-1$ and that k is even.

(1) For continuity of the wave function,

$$G_k(w) + H_k(w) = G_{k-1}(w) + H_{k-1}(w).$$

(2) For discontinuity of the normal derivative,

$$2ik \sin w\{G_k(w) - H_k(w) - G_{k-1}(w) + H_{k-1}(w)\}$$
$$= g_k\{G_k(w) + H_k(w) + G_{k-1}(w) + H_{k-1}(w)\}.$$

We put these equations in the form

$$\begin{aligned} H_{k-1}(w) &= R_k(w)G_{k-1}(w) + T_k(w)G_k(w), \\ H_k(w) &= T_k(w)G_{k-1}(w) + R_k(w)G_k(w), \end{aligned} \tag{4.2}$$

together with a corresponding set for the odd boundaries

$$\begin{aligned} G_k(w+\alpha) &= R_{k+1}(w)H_k(w-\alpha) \\ &\quad + T_{k+1}(w)H_{k+1}(w-\alpha), \\ G_{k+1}(w-\alpha) &= T_{k+1}(w)H_k(w-\alpha) \\ &\quad + R_{k+1}(w)H_{k+1}(w-\alpha). \end{aligned} \tag{4.3}$$

The T_k and R_k are, respectively, the delta function transmission and reflection coefficients,

$$\begin{aligned} T_k &= 2ik \sin w/(2ik \sin w + g_k), \\ R_k &= -g_k/(2ik \sin w + g_k). \end{aligned} \tag{4.4}$$

Thus the G's and H's that satisfy the delta function boundary conditions are constrained to satisfy a set of first-order matrix difference equations. The $G_k(w)$ and $H_k(w-\alpha)$ each form the elements of a column vector of N entries. From this point forward the index k on G and H will be understood.

Here G and H are to be interpreted as column vectors with as many entries as there are regions in the state space. All of the equations (4.2) and (4.3) may be written

$$G(w+\alpha) = M_o H(w-\alpha), \quad H(w) = M_e G(w), \tag{4.5}$$

where the M_o and M_e are $N \times N$ matrices (o for odd boundaries, e for even boundaries). The M_o is made up of m 2×2 matrices spanning the diagonal with the upper left-hand corner of each 2×2 in the odd locations (1,1), (3,3), (5,5),..., etc. Here M_e consists of m 2×2 matrices spanning the diagonal in the even locations, upper left corner in (2,2), (4,4), etc.

Equations (4.5) represent coupled matrix first difference equations: coupled, because G depends on H and vice versa; first difference equations, because G and H are each shifted by one unit of α from the right-hand to the left-hand side of the equations; matrix difference equations, because M_o and M_e are matrices of rank N.

Because of the unitarity of the transmission and reflection coefficients the matrices M_o and M_e are unitary.

This algebraic structure is not dependent upon the delta function potentials. Any unitary functional form for the reflection and transmission coefficients could be substituted, and the corresponding solution to the difference equations could be interpreted as a particle problem. It will be shown in subsequent work that this matrix of transmission and reflection coefficients, which appears here as the coefficient matrix in a difference equation, is the "transfer matrix" of a lattice problem. In this way correspondences between particle problems and lattice problems may be identified.

B. Symmetries of the difference equations

We seek a solution to the matrix difference equations (4.5). By analogy with differential equations we shall exploit the symmetries of the matrix coefficients of these difference equations to reduce the algebra required for their solution.

1. Periodicity

The matricies M_o and M_e of (4.5) are unitary $N \times N$ matrices whose elements are rational functions of e^{iw}, and hence they are periodic with period 2π in w.

2. Unitarity

Unitarity has a special meaning in this context. The elements of these matrices are analytic functions. Unitary matrices are inverted by conjugate transposition, and ordinarily the complex conjugate of an analytic function is not an analytic function. If unitarity is to have an algebraic meaning, conjugation must transform analytic functions into analytic functions; that is, conjugation must be a conformal transformation. Inspection of the matrix elements (4.4) shows that either of the transformations

$$e^{iw} \to e^{-iw}, \quad z \to z^{-1}, \quad e^{iw} \to e^{i(w+\pi)}, \quad z \to -z,$$

carry analytic functions into analytic functions, and invert the symmetric matrices M_o and M_e. We may choose either of these transformations to represent conjugation. The elements of the matrices, the elements of their inverses, and the elements of their complex conjugates will then be analytic functions of $z = e^{iw}$.

There are an infinity of meanings that can be given to complex conjucation in the w plane, due to the multiplicity of values $z \to e^{2in\pi}z$, or the periodicity in w. We may choose any of these meanings at our conveneicne.

3. *Time reversal*

From the form of the matrix elements given in (4.2)–(4.4), the matrices M_o and M_e may be seen to have the properties

$$\begin{aligned} M_o(-w) &= M_o(w+\pi) = M_o^{-1}(w) = M_o^*(w), \\ M_e(-w) &= M_e(w+\pi) = M_e^{-1}(w) = M_e^*(w). \end{aligned} \tag{4.6}$$

Again, these properties appear to be a consequence of delta function transmission and reflection coefficients. In fact they are properties of two-particle one-dimensional transmission and reflection coefficients. They arise from time reversal symmetry and certain constraints on the two-particle interaction; these constraints are satisfied for a wide range of interactions.

C. Symmetries of the solutions to the difference equations

Symmetries of the difference equations allow us to generate solutions by applying the transformations associated with those symmetries. Since the equations are difference equations the associated transformations are discrete.

1. *The translations*

Given any solution to the equations (4.5) an infinite class of further solutions may be generated by repeated translation by 2π in w. We suppose G and H to be solutions of (4.5), and n an integer, then G_n and H_n are also solutions, where

$$G_n = G(2n\pi + w), \quad H_n = H(2n\pi + w), \tag{4.7}$$

because the matrices M_o and M_e are period 2π in w.

2. *The involutions*

The combination of unitarity and time reversal that gives the identities (4.6) leads to a class of involutions or reflexive transformations that transform solutions into solutions. Suppose G and H to be solutions of (4.5). Let

$$G'_n(w) = H(2n\pi - w), \quad H'_n(w) = G(2n\pi - w), \tag{4.8}$$

where n is any integer. This transformation is an involution; if repeated the original G and H are recovered.

The functions G' and H' satisfy

$$H'(2n\pi - w - \alpha) = M_o G'(2n\pi - w + \alpha),$$
$$G'(2n\pi - w) = M_e H'(2n\pi - w).$$

We substitute $w \to 2n\pi - w$ and use the properties (4.6) to obtain

$$G'(w+\alpha) = M_o H'(w-\alpha), \quad H'(w) = M_e G'(w),$$

which are Eqs. (4.5). Thus any particular solution to (4.5) leads to an infinite class of solutions generated by these involutions.

3. *The general solution*

The general solution is an arbitrary linear combination of all possible translations and involutions. The general G and H may therefore be written

$$\begin{aligned} G(w) &= \sum a_n G_n + \sum b_l G'_l, \\ H(w) &= \sum a_n H_n + \sum b_l H'_l, \end{aligned} \tag{4.9}$$

where $-\infty < n < \infty$, $-\infty < l < \infty$, and the a_n, b_l are constants independent of w.

These symmetry properties are sufficient to assure that an algebraic solution to the finite difference equations exists. Further symmetries that come about due to pairwise interactions and the functional form of the transmission and reflection coefficients will be discussed in subsequent work.

4. *The difference equations*

The symmetries of periodicity, unitarity, and time reversibility make it possible to represent the general solution as a linear combination of all translations and involutions of a particular solution. We therefore seek a particular solution to the matrix difference equations (4.5),

$$G(w+\alpha) = M_o H(w-\alpha), \quad H(w) = M_e G(w).$$

The matrices M_o and M_e are unitary matrices. Their elements are rational functions of $z = e^{iw}$. Some economy of presentation is effected if we write the difference equation as a function of z. Recall that $\alpha = 2\pi/N = \pi/m$, and let

$$\omega^N = \omega^{2\pi/\alpha} = 1.$$

The difference equations then become

$$G(\omega z) = M_o(z)H(\omega^{-1}z), \quad H(z) = M_e(z)G(z).$$

These two forms of the difference equations are equivalent, but in context one is often preferred over the other. In what follows we shift freely from one representation to the other.

V. THE ALGEBRAIC SOLUTION OF MATRIX DIFFERENCE EQUATIONS

A. The matrix Riemann–Hilbert functional equation

We iterate Eqs. (4.5) to produce a matrix Riemann–Hilbert functional equation

$$\begin{aligned} &G(w+2\pi) \\ &\quad = M_o(w+2(N-1)\alpha)M_e(w+(N-1)\alpha) \\ &\qquad \times \cdots \times M_e(w)G(w) = N_e(w)g(w), \\ &H(w+2\pi-\alpha) \\ &\quad = M_e(w+2(N-1)\alpha)M_o(w+(N-1)\alpha) \\ &\qquad \times \cdots \times M_o(w)H(w-\alpha) = N_o(w-\alpha)H(w-\alpha). \end{aligned} \tag{5.1}$$

With the change of variable $z = e^{iw}$ these equations read

$$G(w^N z) = N_e G(z),$$
$$H(w^{N-1}z) = N_o(w^{-1}z)H(w^{-1}z),$$

where $\omega^N = 1$, and

$$N_e = M_o(\omega^{N-1}z)M_e(\omega^{N-2}z)\cdots M_e(z),$$
$$N_o = M_e(\omega^N z)M_o(\omega^{N-1}z)\cdots M_o(\omega z). \tag{5.2}$$

Equations (5.1) are matrix Riemann–Hilbert functional equations. They differ from the functional equations of Refs. 4 and 5 by having matrix rather than scalar difference coefficients. Every solution of (4.5) satisfies (5.1). Thus every solution to the difference equations satisfies a matrix Riemann–Hilbert functional equation.

We shall relate particular solutions of matrix difference equations of this type to the eigenvectors of the coefficient matrices. The eigenvectors of these matrices have components that are analytic functions of $z = e^{iw}$. We shall undertake in a separate work a justification and discussion of some of the analytic properties of the eigenvalues and eigenvectors of matrices, particularly as they may be deduced by computer assisted algebraic computation. In the interest of continuity, readability, and brevity, we will proceed here by stating the results.

B. The algebraic properties of the matrices N_o and N_e

1. Eigenvalues

The matrices N_μ, where μ is a collective index for either o or e, are matrices whose elements are rational functions of $z = e^{iw}$. They are therefore period 2π in w. They have further properties that arise from the unitarity and time reversibility of the matrices M_μ.

Note, from (5.2), the identities

$$N_e(\omega z) = M_o N_o(\omega^{-1}z)M_o^{-1},$$
$$N_o(z) = M_e N_e(z)M_e^{-1}. \tag{5.3}$$

From the second of these identities we see that N_o and N_e are similar, and thus have the same characteristic equation and the same eigenvalues.

(1) The characteristic equation of N_μ is

$$P(\lambda,z) = \mathrm{Det}(N_\mu - \lambda I) = 0.$$

It is usual to regard this equation as defining λ as a function of z on a Riemann surface of N sheets ($N = 2m$ is the rank of N_μ). In this multiple value view we parametrize λ as an N-valued function of z.

The algebraic relation between λ and z is independent of parametrization. It is equally valid to parametrize λ and z with a single parameter $w = -i\ln z$. In this view both λ and z are many valued functions of w. By virtue of the algebraic relation between λ and z both are simple automorphic[7] functions of w, and invariant with respect to the same group of fractional linear transformations.

(2) From the first identity of (5.3),

$$P(\lambda,\omega z) = P(\lambda,\omega^{-1}z) = 0, \quad \lambda(\omega z) = \lambda(\omega^{-1}z),$$

and the coefficients of the characteristic equation of N_μ are functions only of z^m, or, what is the same thing, the eigenvalues of N_μ are periodic with period 2α in w.

From the time reversal–unitarity properties (4.5), coupled with the definition (5.2) we have another identity

$$N_0^{-1}(z^{-1}) = N_e(z), \tag{5.4}$$

which implies the following.

(3) The eigenvalues of N_μ are unimodular,

$$P(\lambda^{-1},z^{-1}) = P(\lambda,z), \quad \lambda(z) = \lambda^{-1}(z^{-1}), \quad \lambda\lambda^* = 1, \tag{5.5}$$

if we interpret conjugation as per Sec. IV B 2.

The characteristic equation of N_μ, $P(\lambda,z) = 0$, generates a field of algebraic functions of z. This field of algebraic functions contains all functions that may be written in the form

$$Q(z) = \sum_{n=0}^{N-1} Q_n(z)\lambda^n.$$

Functions $Q(z)$ in this field satisfy an algebraic equation with coefficients that are rational functions of z. The degree of this equation is $\leqslant N$.

2. Eigenvectors

The identities (5.3) are basis independent and apply to the matrices N_μ in diagonal form. Consider the algebraically defined matrix

$$\Lambda_\mu(\lambda,z) = -(N_\mu - \lambda I)^{-1}P(\lambda,z)\left(\frac{\partial P}{\partial\lambda}\right)^{-1}.$$
$$= \text{matrix of cofactors of}(N_\mu - \lambda I)$$
$$\times\left\{\frac{\partial}{\partial\lambda}\mathrm{Det}(N_\mu - \lambda I)\right\}^{-1},$$

where the complex variables λ and z are related so that

$$\mathrm{Det}(N_\mu - \lambda I) = P(\lambda,z) = 0.$$

Here Λ_μ has the following properties:

(1) Λ_μ is a commutative eigenvector matrix of N_μ,

$$N_\mu\Lambda_\mu = \Lambda_\mu N_\mu = \lambda\Lambda_\mu. \tag{5.6}$$

(2) The eigenvectors in Λ_μ are normalized

$$\mathrm{Tr}\ \Lambda_\mu = 1.$$

(3) The relations (5.3) imply

$$\Lambda_e(\omega z) = M_o\Lambda_o(\omega^{-1}z)M_0^{-1},$$
$$\Lambda_o(z) = M_e\Lambda_e(z)M_e^{-1}. \tag{5.7}$$

(4) Unitarity and time reversibility (4.5) imply, from (5.4),

$$\Lambda_o(z^{-1}) = \Lambda_e(z), \quad \Lambda_o(-w) = \Lambda_e(w). \tag{5.8}$$

(4) These eigenvector matricies are orthonormal projection matrices:

$$\text{if}\ \lambda = \beta, \quad \Lambda_\mu(\lambda,z)\Lambda_\mu(\beta,z) = 0;$$
$$\text{if}\ \lambda = \beta, \quad \Lambda_\mu^2(\lambda,z) = \Lambda_\mu(\lambda,z).$$

(5) The elements of Λ_μ are algebraic functions of z. These elements therefore take on all values if they are not constant. Of particular interest in what follows will be the places in the complex z plane where they are infinite and where they are zero.

(a) Some, but not all, of the elements of Λ_μ must have zeros at the zeros of det N_μ and det N_μ^{-1}.

(b) Infinities of the elements of Λ_μ may only occur at the zeros of the algebraic function

$$\eta = \frac{\partial P}{\partial \lambda}.$$

These infinities occur at the simultaneous zeros of $P(\lambda,z)$ and $\partial P/\partial\lambda$. They are therefore branch points and not poles.

The matrix Λ_μ is the algebraically determined analog of a normalized eigenvector. The properties stated here are the algebraic analog of the familiar properties of the eigenvectors and eigenvalues of a unitary matrix.

(6) Here Λ_μ is a matrix that is the outer (or tensor) product of a normalized eigenvector with its adjoint (conjugate transpose). This representation of Λ_μ is a convenient and economical way to account for the properties Λ_μ.

Let v be an m-component column vector and satisfy

$$\begin{aligned} &N_\mu v_\mu = \lambda v_\mu, \quad v_\mu^{*T} N_\mu = \lambda v_\mu^{*T}, \\ &v_\mu^{*T} v_\mu = 1, \quad \Lambda_\mu = v_\mu \times v_\mu^{*T}. \end{aligned} \tag{5.9}$$

These properties of v_μ may be seen to account for all of the properties of Λ_μ.

The elements of Λ_μ are said to be in the field of algebraic functions generated by $P(\lambda,z) = 0$. One consequence of this property is that we may write an algebraic equation satisfied by each of these functions with z as the independent variable. When it is possible to write such an algebraic equation for a function of z, we say that the function is algebraically determined.

The individual components of the normalized eigenvectors are not algebraically determined. Any set of components may be multiplied by an arbitrary function of z such that

$$\rho(z)\rho^*(z) = \rho(z)\rho(-z) = \rho(z)\rho(z^{-1}) = 1,$$

which may be seen to leave all of the elements of Λ unaffected. This independence of eigenvectors on eigenvector phase is a symmetry, often called *gauge symmetry*.

(7) The relations (5.7)–(5.9) imply

$$p_o v_e(\omega z) = M_o v_o(\omega^{-1} z), \quad p_e v_o(z) = M_e v_e(z), \tag{5.10}$$

where $p_e(z) p_e(z^{-1}) = p_0(z) p_0(z^{-1}) = 1$. In addition there is an identity that is required by (5.6),

$$\lambda = \prod_{n=0}^{m-1} p_e(\omega^{2n} z) \prod_{n=0}^{m-1} p_o(\omega^{2n+1} z). \tag{5.11}$$

C. Eigenvector solutions to the matrix difference equations

The functions p_μ are not generally algebraic functions. They need not be in any algebraic function field. The arbitrary nature of p_μ arises from the symmetry of independence of eigenvector phase, which is from gauge symmetry. It is always possible to choose a basis eigenvector phase, or gauge, such that p_μ^2 is in the field of algebraic functions generated by the characteristic equation of N_μ. The details of how to do this will be presented elsewhere. We shall proceed here by assuming that this has been done. When this has been done we shall say that the eigenvectors are in standard form.

Let $G = s_e v_e$, $H = s_o v_o$, and Eqs. (5.10) become

$$\{p_o s_e^{-1}(\omega z) s_o(\omega^{-1} z)\} G(\omega z) = M_o H(\omega^{-1} z),$$

$$\{p_e s_o^{-1}(z) s_e(z)\} H(z) = M_e G(z).$$

The matrix difference equations are solved if we find analytic functions s_o and s_e such that

$$s_{ee}(\omega z) = p_o s_o(\omega^{-1} z), \quad s_o(z) = p_e s_e(z). \tag{5.12}$$

The matrix difference equations are solved provided that a proper solution to these difference equations exists. A proper solution is a solution consistent with the analyticity constraints of the Sommerfeld ansatz. Every solution to either of Eqs. (5.12) satisfies a scalar Riemann–Hilbert functional equation of the form

$$s(w + 2\pi) = \lambda(w) s(w), \tag{5.13}$$

where $\lambda(w)$ is an algebraic function of $z = e^{iw}$. An alternative form of the equation is

$$s(\omega^N z) = \lambda(z) s(z). \tag{5.13$'$}$$

It is possible to show that if we are given a particular solution to the Riemann–Hilbert functional equation, we can construct a solution to (5.12). We find it convenient to defer this argument to Sec. V E.

D. The solution of scalar Riemann-Hilbert functional difference equations

The difference coefficient in the scalar Riemann–Hilbert functional equation $\lambda(w)$ is an algebraic function of $z = e^{iw}$; λ and z are related by the characteristic equation of N_μ, a rational function of the form

$$\det(N_\mu - \lambda I) = \sum \sum A_{ij} \lambda^i z^j = 0, \tag{5.14}$$

where A_{ij} are constants.

The degree of λ is $2m$, the rank of N_μ. The degree of z depends upon the specific form of the transmission and reflection coefficients for the two-body problem. From the periodicity of the characteristic equation z appears only as z^m.

Functional equations of this type have been solved by exploiting some special form of the characteristic equation. In the most frequently occurring case the Bethe ansatz is satisfied, z does not appear, and all the roots are $\lambda = 1$. The impenetrable case of Ref. 4 is a special degenerate case where N_μ is of rank 1, and λ is a rational function of z. The case analyzed by Gaudin and Derrida[5] will be shown in a later work to correspond to a case where $m = 3$, the characteristic equation is reducible to a quadratic in λ, and in this quadratic z appears only to powers 12, 6, and 0.

The function $\lambda(w)$ is an automorphic function, invariant with respect to a group of fractional linear transformations T. Let Q be a subgroup of those transformations such that $Q(w) = w + 2\pi$. We phrase the Riemann–Hilbert functional equation as

$$s(Q(w)) = \lambda(w) s(w).$$

Since λ is automorphic with respect to T, $\lambda(T(w)) = \lambda(w)$ is true for arbitrary w. It is true then, that

$$s(QT) = \lambda s(T).$$

Suppose that $s(T(w))$ is a solution to (5.13). Then

$$s(TQ) = \lambda s(T).$$

By functionally inverting the analytic function s we see that

$$QT = TQ = s^{-1}(\lambda s),$$

and therefore Q and T commute. It is not difficult to show that the only transformation that commutes with a translation is another translation. If there are more than two linearly independent translations the group is continuous, and not allowed. Thus if solutions transform into solutions under all of the transformations of the group of automorphisms of λ, the group is singly or doubly periodic.

There is no reason to assume that the characteristic equation of N_μ will be such that all of the automorphisms of λ are periods. We shall show in a later work that the cases previously cited are the only cases where this will be true.

It is well known that any singly or doubly periodic function may always be represented as a ratio of theta functions, or as an elliptic function.[2,3,7] An important consequence of this restricted functional form is that the solution to the Riemann–Hilbert functional equation may be written as a ratio of infinite products of linear factors; through the theta function representation in the doubly periodic case[5]; through a polynomial representation in the singly periodic case.[4]

The solutions generated in either the singly or doubly periodic case are in the field of functions generated by the characteristic equation of N_μ. They possess all of the automorphism of this field, and no others. In the general case this will not be true. The general restriction on the solution of the Riemann–Hilbert functional equation is that it be analytic, to satisfy the constraints of the Sommerfeld ansatz. It need not belong to any particular function field or be automorphic with respect to any particular function group.

Particular analytic solutions to (5.13) are not difficult to construct. Again, make an ansatz and look for a solution of the form

$$s=\lambda^{\nu}. \tag{5.15}$$

This is a solution if ν satisfies the difference equation

$$\nu(w+2\pi)-\nu(w)=1,$$

which has the particular solution

$$\nu=w/2\pi.$$

The general solution is obtained by adding a solution to the homogeneous equation

$$\nu=w/2\pi+\phi(w),$$

where $\phi(w+2\pi)=\phi(w)$.

At each pole or zero of λ in the w plane a new branch of s is encountered. If λ is not singly or doubly periodic these branches cannot be removed and the solution exists on a Riemann surface of infinite genus, which means that the solution is restricted to a particular sheet of the function λ.

This solution is analytic and may be expanded in a power series with a finite radius of convergence at any ordinary point of the surface of $\lambda(w)$. Almost every point is an ordinary point; i.e., the poles, zeros, and branch points of λ and s are isolated.

The functional equation for s leads to poles and zeros of increasing multiplicity the greater the range of re(w), as may be seen from the w dependence of ν. This is characteristic of the solution of Riemann–Hilbert functional equations.[4,5]

The solution to the difference equation therefore requires that we pick a particular sheet of the algebraic functions λ and remain on that sheet for all path integrations. This restriction is to be maintained in all that follows.

E. Unimodular solutions to the scalar Riemann–Hilbert functional equation

In Sec. IV we found the operation of complex conjugation was identified with either of a pair of fractional linear transformations $z\rightarrow z^{-1}$ or $z\rightarrow -z$. Unimodular functions are functions (like λ and p_μ) whose reciprocal is their conjugate.

Since complex conjugation is an operation that carries analytic functions into analytic functions, it preserves functional relations and we may write from (5.12)

$$s_e{}^*(\omega z)=p_o{}^*s_o{}^*(\omega^{-1}z),\quad s_o{}^*(z)=p_e{}^*s_e{}^*(z). \tag{5.16}$$

Multiply the left- and right-hand sides by the left- and right-hand sides of (5.12), respectively, and note that $p_\mu p_\mu{}^*=1$. Thus

$$s_e{}^*(\omega z)s_e(\omega z)=s_o{}^*(\omega^{-1}z)s_o(\omega^{-1}z),$$
$$s_o{}^*(z)s_o(z)=s_e{}^*(z)s_e(z).$$

This shows that function $s_\mu(z)s_\mu{}^*(z)$ is a function only of z^m and is therefore periodic with period 2α in w.

Since every solution is only determined up to a multiple period 2α in w, we may, without loss of generality, choose the functions s_μ to be unimodular, that is

$$s_es_e{}^*=s_os_o{}^*=1.$$

We replace z by z^{-1} in (5.16) and rearrange to obtain

$$s_e{}^*(\omega z^{-1})=p_o{}^*s_e{}^*(\omega^{-1}z^{-}1),\quad s_e{}^*(z)=p_e{}^*s_o{}^*(z). \tag{5.17}$$

Again, respectively, multiplying left- and right-hand sides by left- and right-hand sides of (5.12),

$$s_e(\omega z)s_o{}^*(\omega z^{-1})=s_o(\omega^{-1}z)s_e{}^*(\omega^{-1}z^{-1}),$$
$$s_o(z)s_e{}^*(z)=s_e(z)s_o{}^*(z). \tag{5.18}$$

The Theorem of Unimodularity: If s_μ is unimodular and

$$F=s_e(z)s_o{}^*(z^{-1}), \tag{5.19}$$

the difference equations imply that

$$FF^*=1\quad\text{and}\quad F(\omega^2z)=F(z).$$

The function F is unimodular, a function only of z^m and period 2α in w.

In the w plane with the particular choice of involution such that $z\rightarrow z^{-1}$ implies $w\rightarrow 2\pi-w$, implies

$$F=s_e(w)s_o{}^*(2\pi-w),$$

which we shall use in satisfying a class of asymptotic boundary conditions.

F. An explicit connection to the Riemann–Hilbert functional equation

To provide an explicit solution to Eqs. (5.12) it is convenient to separate them by the substitution

$$s_e(z)=B(z)A(\omega z),\quad s_o(z)=B(\omega^2z)A(\omega z),$$

which, upon substitution in (5.12), yields the second difference equations

$$B(\omega^2 z) = p_e B(z), \quad A(\omega^2 z) = p_o A(z).$$

In these second difference equations only the second difference appears, so that we may solve them by the first difference techniques. Let us seek a particular solution that is unimodular.

We let

$$B(z) = B_e(z)\lambda^{w/2\pi}, \quad A(z) = A_o(z)\lambda^{w/2\pi},$$

and

$$B_e(\omega^2 z) = p_e \lambda e^{-1/m} B_e(z),$$
$$A_o(\omega^2 z) = p_o \lambda_o^{-1/m} A_o(z),$$

where

$$\lambda_e = \prod_{n=0}^{m-1} p_e(\omega^{2n} z), \quad \lambda_o = \prod_{n=0}^{m-1} p_o(\omega^{2n} z).$$

The expressions

$$B_e = \left\{ \prod_{n=1}^{m} p_e{}^n(\omega^{2(n-1)} z) \right\}^{1/m},$$
$$A_o = \left\{ \prod_{n=1}^{m} p_o{}^n(\omega^{2(n-1)} z) \right\}^{1/m},$$

are unimodular solutions to (5.12). The ambiguity that arises in the taking of the mth root is equivalent to multiplication by a unimodular function period 2α, the function F of the previous section. This ambiguity is to be resolved with the application of boundary conditions.

G. Summary of the solution to the matrix difference equations

A particular solution to the coupled first difference equations (4.4) is

$$G(w) = s_e(w) v_e(w), \quad H(w) = s_o(w) v_o(w),$$

where $v_e(w)$ and $v_o(w)$ are standard form normalized eigenvectors of $N_e(w)$ and $N_o(w)$, respectively, and s_μ is a particular solution to

$$s_e(\omega z) = p_o s_o(\omega^{-1} z), \quad s_o(x) = p_e s_e(z).$$

where$p_e^2(w)$ and$p_o^2(w)$ are algebraic functions in the field of functions generated by the characteristic equation of N_μ.

A unimodular solution $s_\mu(w)$ is constructed from p_e, p_o, and $s(w)$, a particular solution to a Riemann–Hilbert functional equation

$$s(w + 2\pi) = \lambda(w) s(w),$$

where λ is a root of the characteristic equation of N_μ.

In order to answer specific questions about the quantum system under investigation, we require that we be able to choose $s(w)$ such that linear combinations of translations of involutions give rise to state functions that satisfy boundary conditions specific to some physical situation.

These boundary conditions are conditions on the state function on some boundary of the state space. In order to satisfy conditions of this type we need to be able to see the behavior of the integrals of the Sommerfeld ansatz in the vicinity of the boundary.

VI. THE INTEGRALS OF THE SOMMERFELD ANSATZ

We have seen how to determine a class of particular solutions to the difference equations generated by the Sommerfeld ansatz. The complete determination of the solution to the partial differential equation requires that a particular solution be modified to satisfy boundary conditions on the state function. We now represent the integrals of the Sommerfeld ansatz for the special case of a unimodular eigenvector solution to the difference equations. We wish to see how these integrals reveal the behavior of the state functions they generate.

A. Involution and translation reduction of the integrals of the Sommerfeld ansatz

In Sec. IV we found that the complete solution to the coupled first difference equations is given by finding a particular solution of the coupled matrix first difference equations for G and H; the remaining linearly independent solutions are generated by repeated application of involution and translation transformations. The general solution is to be written as a sum over translations and involutions of a particular solution, as in (4.7)–(4.9). These forms are then substituted into the basic integral representation of the Sommerfeld ansatz (4.1). All of the integrals are to be computed on the basic contour of Fig. 2.

The complete state function is

$$\Psi = \sum_{n=-\infty}^{\infty} a_n \Psi_n + \sum_{l=-\infty}^{\infty} b_l \Psi'_l,$$

where

$$\Psi_n = \int_C [G(w + \theta + 2n\pi) + H(2n\pi + w - \theta]$$
$$\times e^{ikr\cos w} dw, \tag{6.1}$$
$$\Psi'_l = \int_C [G(-w + \theta + 2l\pi) + H(2l\pi - w - \theta)]$$
$$\times e^{ikr\cos w}\, dw.$$

We substitute the unimodular eigenvector solutions of Sec. V for G and H, making repeated use of the fact that the s_μ satisfy the Riemann–Hilbert functional equation (5.13). The integral Ψ_n is evaluated on the basic contour of Fig. 2,

$$\Psi_n = \int_C [\lambda^n(w + \theta) s_e(w + \theta) v_e(w + \theta)$$
$$+ \lambda^n(w - \theta) s_o(w - \theta) v_o(w - \theta)]$$
$$\times e^{ikr\cos w}\, dw; \tag{6.2}$$

and Ψ'_l is evaluated on the lower half-plane image of the basic contour; $C'(w) = C(-w)$,

$$\Psi'_l = \int_{C'} [\lambda^l(w + \theta) s_e(w + \theta) v_e(w + \theta)$$
$$+ \lambda^l(w - \theta) S_o(w - \theta) v_o(w - \theta)]$$
$$\times e^{ikr\cos w}\, dw.$$

Every choice of a_n and b_l leads to a linearly independent solution. Each of the integrals is a linearly independent solution. These two linearly independent solutions for the same n, Ψ_n, Ψ'_n, have the same integrand, but differ in the placement of the contour in the complex w plane.

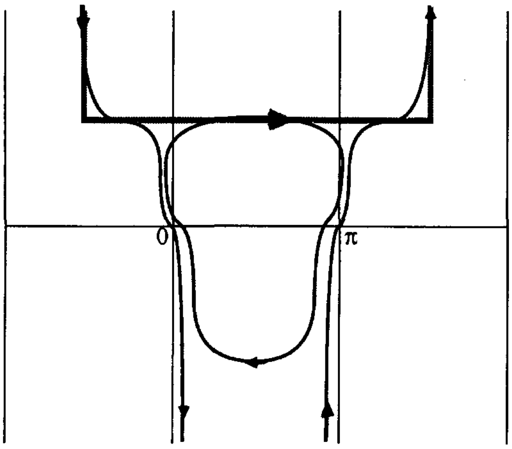

FIG. 4. The basic contour and its deformation.

B. Deformation of the basic contour

We now deform the basic contour as shown in Fig. 4. Only two types of paths remain: open contours that begin and end at infinity and pass through a single stationary phase or steepest descent point; closed contours that enclose singularities, either poles or branch lines. Each of these types of path give a characteristic contribution to the state function. The lower half-plane image of the basic contour is deformed in the same way.

1. Contribution from an open contour passing through the steepest descent point at $w=0$

At every point on the path the integrand is of the form $e^{ikr\cos w}$, where $\cos w$ has a positive real part and therefore the phase is increasing in the direction of positive r as time increases. All contributions on this path are waves diverging from the center of mass of the three-particle system. Thus any contour that passes through the steepest descent point at any even multiple of π contributes a state function made up of only outgoing waves.

2. Contribution from an open contour passing through the steepest descent point at $w=\pi$

At every point along the path the contribution is of the form $e^{ikr\cos w}$, where $\cos w$ has a negative real part and therefore the phase is increasing in the direction of negative r as time increases. All contributions on this path are waves converging upon the center of mass of the three-particle system. Thus any contour that passes through the steepest descent point at any odd multiple of π contributes a state function made up of only incoming waves.

3. Contribution from a closed contour

The closed path surrounds an area of the complex w plane, exclusive of the steepest descent points at $w=0$, $w=\pi$. Inside this contour there are, in general, both poles and branch points connected by branch lines that cannot be crossed by the contour.

Poles of s_μ within a closed contour will contribute plane waves. Plane waves are eigenstates of a two-body system. Their presence is signaled by a pole of det N_μ.

Infinities of the components of eigenvectors appear at the branch line paths within the closed portions of C and C' the line path integral is of the form

$$\int_a^b F(w)e^{ikr\cos(w-\theta)}\,dw,$$

where a and b are branch points. The state function that arises from a line path integral is not interpretable as either of the above types of state function. The paths connecting the branch points are not allowed to approach the steepest descent point, so that the integrand is typically rapidly oscillating. An exception to this typical behavior occurs when $k^2\leqslant 0$ and the branch points are real.

C. Interference of Ψ_n and Ψ'_l

We wish to emphasize that although these path integrals generate a complete set of linearly independent state functions, the state functions are not orthogonal. The probabilities generated from the absolute squares of these state functions will not be probabilities of independent events.

The bewildering array of linearly independent solutions arises from the fact that a bewildering array of possible physical situations are consistent with this formulation. The Sommerfeld ansatz has provided an algebraic solution to any three-particle problem where the particles interact locally (i.e., when their coordinates are identical) within a bounded region of state space that includes the center of mass. This bounded region can be finite, as in periodic boundary conditions. The outside of this bounded region is not a part of the state space of the Sommerfeld ansatz, and the freedom to accommodate any sort of interaction outside that region accounts for the wide range of possible solutions.

VII. THE ASYMPTOTIC SOLUTION

A. Asymptotic boundary conditions

The problem of fixing G and H for arbitrary boundary conditions is formidable. We concentrate our efforts here upon finding an asymptotic solution. This asymptotic solution is to exist in the wedge shaped region bounded by the limit $r\to\infty$, and the two lines $\theta=0$ and $\theta=\alpha$ in each region of the state space.

A simplification of an asymptotic solution is that the spectrum is known. The only amplitudes which survive as $r\to\infty$ correspond to energies that belong to a continuum beginning at the binding energy of the lowest two-particle bound state. We will assume that in at least one channel the interaction is attractive and a two-particle bound state exists. This assumption is not required to satisfy any algebraic constraint. It is adopted so that some normalizable state may be analytically continued from $k^2>0$ to $k^2<0$.

These scattering, or zero density, boundary conditions are not all-inclusive. The asymptotic result may always be recovered from a wave packet argument as the zero density limit of any finite density solution, but it is not clear what features of a finite density solution may be derived from the asymptotic limit.

We have seen that all solutions may be derived from linear combinations of translations and involutions of any particular solution to the difference equations. It will be convenient to base our discussion upon the eigenvector particular solutions of Sec. V.

In order that a proper normalizable solution exist within an infinite wedge the wave function must satisfy certain conditions.

(1) The wave function must be regular at $r = 0$, which, for eigenvector solutions, requires that $s_\mu(w)$ be bounded as $w \rightarrow i\infty$.

(2) The wave function must not increase exponentially in any direction within the wedge.

(3) We shall make it a condition of our asymptotic solution that only two-particle bound pair plane wave states exist in the asymptotic limit.

In the asymptotic limit open contours contribute only at the steepest descent point, and give rise to what we shall call "free waves." These free waves are either incoming which converge upon, or outgoing which diverge from the center of mass of the three-particle system. Condition (3) does not allow contributions from the branch line paths to arise asymptotically. In our analysis this condition is met by open contours blocking the branch line paths from the steepest descent point. Only open paths are allowed to pass through these points.

Condition (3) forbids a plane wave asymptotic solution unless the plane wave is associated with a two-particle bound pair. These bound pair plane waves come from poles within the closed contours.

B. The unimodular eigenvector particular solutions

The unimodular eigenvector particular solutions automatically satisfy condition (1). We rewrite the integral representations for these solutions as

$$\Psi_n = \Gamma_e^n(\theta) + \Gamma_o^n(\theta), \quad \Psi_n' = \Gamma_e'^n(\theta) + \Gamma_o'^n(\theta), \tag{7.1}$$

where

$$\Gamma_e^n(\theta) = \int_C \lambda^n(w) s_e(w) v_e(w) e^{ikr\cos(w-\theta)}\, dw,$$

$$\Gamma_o^n(\theta) = \int_C \lambda^n(w) s_o(w) v_o(w) e^{ikr\cos(w+\theta)}\, dw,$$

Γ' is the lower half-plane image path integral, i.e, $\Gamma \rightarrow \Gamma'$ if $C \rightarrow C'$.

The unimodularity theorem (5.17)–(5.19) the property (5.8) that $v_e(\theta) = v_o(-\theta)$, and the aymptotic boundary conditions allow us to represent Γ_o and Γ_e in a single integral representation.

We choose the functions s_μ to be unimodular, so that

$$s_e s_e^* = s_o s_o^* = 1.$$

and

$$F = s_e(w) s_o^*(2\pi - w).$$

The difference equations imply that

$$FF^* = 1 \quad \text{and} \quad F(w + 2\alpha) = F(w),$$

the function F is unimodular and of period 2α. Conditions (2) and (3) imply that F has poles only where λ or λ^{-1} is inifinite. These conditions limit F to the form

$$F = e^{i\psi}\lambda^\delta,$$

where δ and ψ are real parameters. These parameters define a particular solution of the difference equations such that

$$s_e(w) = e^{i\psi}\lambda^\delta s_o(2\pi - w). \tag{7.2}$$

By using the properties of the function λ, and the fact that s_μ satisfies the Riemann–Hilbert functional equation this relation yields the following useful relations:

$$s_o(-w) = e^{-i\psi}\lambda^{-\delta+1} s_e(w),$$

$$s_o(\pi - w) = e^{-i\psi}\lambda^\delta s_e(\pi + w),$$

which, together with the properties above, may be substituted into the integrals (7.1) to obtain a relation among the Γ's:

$$\Gamma_o^n(\theta) = -e^{i\psi}\Gamma_e'^{-n-\delta+1}.$$

This leads to a representation of the state function

$$\Psi_n = \Gamma^n(\theta) - e^{-i\psi}\Gamma'^{-n-\delta+1}(\theta),$$

$$\Psi_n' = \Gamma'^n(\theta) - e^{-i\psi}\Gamma^{-n-\delta+1}(\theta),$$

where

$$\Gamma^n = \int_C \lambda^n(w) s_e(w) v_e(w) e^{ikr\cos(w-\theta)}\, dw,$$

Γ' is the corresponding lower half-plane integral.

We shall now demonstrate that probability fluxes depend upon $s_\mu s_\mu^*$, which is constant, and upon the real parameters ψ and δ. Thus, through the choice of unimodular eigenvector particular solution to the difference equations, we avoid the full complexities of the solution to a Riemann–Hilbert functional equation.

C. Asymptotic unimodular eigenvector solutions

1. Amplitudes of free waves

We deform the basic contour for the integrals Γ as discussed in Sec. VI, and as shown in Fig. 4. The integrals Γ, Γ′ are evaluated asymptotically. The open paths contribute only at the steepest descent points, $w = 0$ and $w = \pi$ for the contour C, $w = 0$ and $w = -\pi$ for the contour C'. The incoming and outgoing asymptotic column vectors are

$$\Psi_{n,\text{out}} = (2\pi/kr)^{1/2} e^{i\pi/4} e^{ikr}\{\lambda^n - e^{-i\psi}\lambda^{-n-\delta+1}\} \times s_e(\theta) v_e(\theta),$$

$$\Psi_{n,\text{in}} = (2\pi/kr)^{1/2} e^{i\pi/4} e^{-ikr}\{\lambda^{-n} - e^{-i\psi}\lambda^{n+\delta}\} \times s_e(\pi + \theta) v_e(\pi + \theta),$$

$$\Psi'_{l,\text{out}} = (2\pi/kr)^{1/2} e^{i\pi/4} e^{ikr}\{\lambda^l - e^{-i\psi}\lambda^{-l-\delta+1}\} \times s_e(\theta) v_e(\theta),$$

$$\Psi'_{l,\text{in}} = (2\pi/kr)^{1/2} e^{i\pi/4} e^{-ikr}\{\lambda^{-l+1} - e^{-i\psi}\lambda^{l+\delta-1}\} \times s_e(\pi + \theta) v_e(\pi + \theta),$$

where $\Psi_{n,\text{out}} = \Psi'_{n,\text{out}}$ for all choices of ψ and δ. Here $\Psi_{n+1,\text{in}} = \Psi'_{n,\text{in}}$ for all choices of ψ and δ.

2. Fluxes of free waves

The incoming and outgoing probability fluxes are formed in accordance with Sec. III B. The outward probability flux normal to a circle of radius r is given by

$$\mathbf{j} = i\left\{\Psi \times \frac{\partial \Psi^{*T}}{\partial r} - \frac{\partial \Psi}{\partial r} \times \Psi^{*T}\right\}\mathbf{n}_r = \frac{P(\theta)\mathbf{n}_r}{r}. \quad (7.3)$$

Here $P(\theta)$ is a probability flux matrix, and is independent of r. For example,

$$P_{n,\text{out}}(\theta) = ir\left(\Psi_{n,\text{out}} \times \frac{\partial \Psi_{n,\text{out}}{}^{*T}}{\partial r} - \frac{\partial \Psi_{n,\text{out}}}{\partial r} \times \Psi_{n,\text{out}}{}^{*T}\right).$$

We substitute the above expressions for the Ψ_n and obtain the probability flux matrices,

$$P_{0,\text{out}} = 4\pi\{2 - e^{-i\psi}\lambda^{-\delta+1} - e^{i\psi}\lambda^{\delta-1}\}\Lambda_e(\theta),$$

$$P_{0,\text{in}} = 4\pi\{2 - e^{i\psi}\lambda^{\delta} - e^{i\psi}\lambda^{-\delta}\}\Lambda_e(\pi+\theta),$$

$$P_{n,\text{out}} = 4\pi\{2 - e^{-i\psi}\lambda^{-2n-\delta+1} - e^{i\psi}\lambda^{2n+\delta-1}\}\Lambda_e(\theta),$$

$$P_{n,\text{in}} = 4\pi\{2 - e^{i\psi}\lambda^{-2n-\delta} - e^{i\psi}\lambda^{2n+\delta}\}\Lambda_e(\pi+q),$$

$P_{n,\text{out}} = P'_{n,\text{out}}, P_{n+1,\text{in}} = P'_{n,\text{in}}$ for all choices of ψ and δ.

The $P(\theta)$ are matrices whose elements are analytic functions of θ. The Riemann–Hilbert functional equation enters only through $s_\mu s_\mu{}^*$, which is constant. In the physical region of θ, where $0<\theta<\alpha$, the diagonal elements of this matrix give probability flux at each value of θ in the region. The off-diagonal elements give the relative phase between the probability fluxes in the various regions.

We have separated the matrices P_0 because they have a special meaning. The number n is like an eigenvalue of angular momentum, and is associated with rotation in the state space of the three-particle system. It corresponds to the minimum length occupied by the three particles in the course of their scattering multiplied by the momentum of the system in the center of mass. The particular case $n = 0$ means that the particles are correlated such that all of the incident flux, either bound or free, is timed so that all of the particles arrive at the center of mass as nearly as possible to the same time. They cannot all arrive at exactly the same time because of the uncertainty principle. The maximum three-particle scattering corresponds to $n = 0$.

A convenient choice of δ is $\delta = \frac{1}{2}$, which makes

$$P_{0,\text{out}}P^{*T}{}_{0,\text{in}} = 4\pi\{2 - e^{-i\psi}\lambda^{1/2} - e^{i\psi}\lambda^{-1/2}\}\Lambda_e(\theta).$$

With this choice of δ the conservation of free flux is manifest for all values of ψ. Not only are the total fluxes equal, the flux in and out at each value of θ are equal. The total flux out in the physical region is given by summing over all regions and integrating over physical θ. The sum over all regions is tr Λ_e, and tr $\Lambda_e = 1$ for all θ. Thus

$$\Phi_{\text{out}}(k^2) = 4\pi\int_0^\alpha (2 - e^{-i\psi}\lambda^{1/2} - e^{i\psi}\lambda^{-1/2})d\theta.$$

This is an integral of the algebraic function $\lambda^{1/2}$, and is irreducible. It is one of the natural functions of this system, and must be evaluated independently.

This choice of δ is convenient for other values of n as well, but the flux conservation is not manifest. There are as many flux conservation integrals as there are independent integrals of the algebraic functions $\lambda^{1/2}$. This number is $3p-3$, where p is the genus of the Riemann surface on which $\lambda^{1/2}$ is uniform.[8]

3. Bound pairs for n=0

The bound pair amplitudes arise only from the presence of poles in $s_e(w)$, because the components of the eigenvectors cannot have poles at the two-particle bound states. From the explicit form of the solution to the functional equation (5.20) we deduce that poles of s_e are colocated in the w plane with a pole of λ or λ^{-1}. Since $\lambda\lambda^{-1} = 1$. The zeros of λ are the poles of λ^{-1} and vice versa. The poles of λ and λ^{-1} are the poles of det N_μ, which are the poles of two-particle reflection and transmission coefficients. Thus the colocation of poles and zeros of s_e with poles of λ or λ^{-1} is consistent with conditions (2) and (3).

We concentrate on the case $n = 0$, where the two integrals to be evaluated are

$$\Gamma^0 = \int_C s_e(w)v_e(w)e^{ikr\cos(w-\theta)}\,dw,$$

$$\Gamma'^0 = \int_C s_e(w)v_e(w)e^{ikr\cos(w-\theta)}\,dw.$$

Consider the asymptotic bound wave arising as the result of the residue of a simple pole at w_0,

$$\Gamma^0 = 2\pi i e^{ikr\cos(w_0-\theta)} \lim_{w\to w_0}(w-w_0)s_e(w)v_e(w)$$
$$= 2\pi i e^{ikr\cos(w_0-\theta)}s_{er}(w_0)v_e(w_0),$$

where

$$s_{er}(w_0) = \lim_{w\to w_0}(w-w_0)s_e(w).$$

the residue of s_e at w_0.

This simple pole must be colocated with a pole of either λ or λ^{-1}. We assume it to be located with a pole of λ, and factor out λ using the difference equation

$$s_{er}(w_0) = \lim_{w\to w_0}(w-w_0)\lambda(w)s_e(w-2\pi),$$

λ has a simple pole at w_0, and $s_e(w_0-2\pi)$ is regular. The residues of s_e at simple poles are therefore directly related to the residues of λ by the relation

$$s_{er}(w_0) = \lambda_r(w_0)s_e(w_0-2\pi),$$

where

$$\lambda_r = \lim_{w\to 0}(w-w_0)\lambda(w).$$

We compute the flux matrix which arises from (7.3). In the case of the bound waves the flux arises from Γ^0, Γ'^0 due to a simple pole in s_e at w and $s_e{}^*$ at w':

$$P_0 = 8\pi^2 ikr(\cos(w-\theta) + \cos(w'^*-\theta))$$
$$\times e^{ikr(\cos(w-\theta) - \cos(w'^*-\theta))}$$
$$\times s_{er}(w)s_{er}{}^*(w')v_e(w) \times v_e{}^*(w')$$
$$= 2ikr(\cos\tfrac{1}{2}(w-w'^*)\cos\tfrac{1}{2}(w+w'^*) - \theta)$$
$$\times e^{-2ikr(\sin\frac{1}{2}(w-w'^*)\sin(\frac{1}{2}(w+w'^*)-\theta))}$$
$$\times(8\pi^2 s_{er}(w)s_{er}{}^*(w')v_e(w) \times v_e{}^*(w')).$$

Conditions (2) and (3) impose further restrictions upon the location of these poles. Let $w = u + i\tau$, $w' = u' + i\tau'$, u, u',τ, and τ' all real. Suppose k to be real and positive, $k^2 > 0$. There are the following choices for the values of u, u', τ,τ', which are consistent with conditions (2) and (3).

(1) $u = u' = 0$, $\tau = \tau' > 0$ contributing

$$= 2\cosh\tau\cos\theta e^{-2kr\sinh\tau\sin\theta} \times (8\pi^2 kr s_{er}(\tau)s_{er}{}^*(\tau)\Lambda_e(\tau)),$$

a matrix that is everywhere exponentially decreasing in the physical region, but that nevertheless carries a net flux at $\theta = 0$ for each value of τ. This flux is calculated by integrating $0 < r\sin\theta < \infty$:

$$\Phi(\theta = 0) = 8\pi^2 \coth\tau s_{er}(i\tau)s_{er}{}^*(i\tau)\Lambda_e(i\tau).$$

The positive sign of this flux indicates that it represents outward probability flow. We use the above rule for the calculation of the residue of s_e at a simple pole, and obtain an expression in terms of the algebraic functions λ,

$$\Phi(\theta = 0) = 8\pi^2 \coth\tau\lambda_r(i\tau)\lambda_r{}^*(i\tau)\Lambda_e(i\tau).$$

There is one such contribution for each value of τ consistent with the binding of the pair on the boundary at $\theta = 0$. The matrix $\Lambda_e(i\tau)$ will be nonzero only for the regions connected at $\theta = 0$ and with a value of τ specific to the strength constant on that boundary.

The same is true for the remaining cases.

(2) $u = u' = \pi + \alpha$, $\tau = \tau' > 0$, contributes incoming fluxes at $\theta = \alpha$:

$$\Phi(\theta = \alpha) = -8\pi^2 \coth\tau\lambda_r{}^{-1}\lambda_r{}^{-1*} \times (\pi + \alpha + i\tau)\Lambda_e(\pi + \alpha + i\tau).$$

(3) $u = u' = \pi$, $\tau = \tau' < 0$, contributes incoming fluxes at $\theta = 0$:

$$\Psi(\theta = 0) = -8\pi^2 \coth\tau\lambda_r(\pi - i\tau)\lambda_r{}^* \times (\pi - i\tau)\Lambda_e(\pi - i\tau).$$

(4) $u = u' = \alpha$, $\tau = \tau' < 0$, contributes outgoing fluxes at $\theta = \alpha$:

$$\Phi(\theta = \alpha) = 8\pi^2 \coth\tau\lambda_r{}^{-1}(\alpha - i\tau)\lambda_r{}^{-1*} \times (\alpha - i\tau)\Lambda_e(\alpha - i\tau).$$

The contributions (1) and (4) are outgoing fluxes, whereas (2) and (3) are incoming. Contributions (1) and (2) are in the upper half-plane, and arise from Γ^0, contributions (3) and (4) are in the lower half-plane and arise from Γ'^0.

In these bound pairs, as in the free fluxes, the conservation of probability flux is manifest. The sum of the fluxes in all channels is the trace $\Lambda_e = 1$, as before. The symmetry of λ,

$$\lambda^* = \lambda(\pi + w),$$

is reflected in the residues. It is therefore true that

$$\Phi_{\text{in}}(\theta = 0) + \Phi_{\text{out}}(\theta = 0) = 0$$

from (1) and (3), whereas

$$\Phi_{\text{in}}(\theta = \alpha) + \Phi_{\text{out}}(\theta = \alpha) = 0$$

from (2) and (4).

A similar argument can be carried out for any value of n. We shall not reproduce it here, but will justify it in subsequent work when we discuss the details of special cases.

D. Conservation of probability flux for $n=0$, $\delta=\frac{1}{2}$

The unimodular eigenvector solutions conserve probability flux for every value of θ in the case $n = 0$, $\delta = \frac{1}{2}$. We analyze the behavior of the flux matrix for all values of k^2, in this case.

For $k^2 > 0$: The outgoing and incoming total free fluxes are, for any value of ψ,

$$\Phi(k^2) = 4\pi\int_0^\alpha (2 - e^{-i\psi}\lambda^{1/2} - e^{i\psi}\lambda^{-1/2})d\theta.$$

This total flux is obviously conserved. From the properties of λ it is not difficult to show that Φ is positive and real. If $\psi = 0$, $\Phi(0) = 0$.

The outgoing and incoming bound pair flux matrices are, for each value of τ consistent with the strength constant on the boundary: along $\theta = 0$,

$$\Phi(\theta = 0) = 8\pi^2 \coth\tau\lambda_r(i\tau)\lambda_r{}^*(i\tau)\Lambda_e(i\tau),$$

$$\Phi(\theta = 0) = -8\pi^2 \coth\tau\lambda_r(\pi - i\tau)\lambda_r{}^*(\pi - i\tau) \times \Lambda_e(\pi - i\tau);$$

along $\theta = \alpha$,

$$\Phi(\theta = \alpha) = -8\pi^2 \coth\tau\lambda_r{}^{-1}\lambda_r{}^{-1*}(\pi + \alpha + i\tau) \times \Lambda_e(\pi + \alpha + i\tau),$$

$$\Phi(\theta = \alpha) = 8\pi^2 \coth\tau\lambda_r{}^{-1}(\alpha - i\tau)\lambda_r{}^{-1*}(\alpha - i\tau) \times \Lambda_e(\alpha - i\tau).$$

The fluxes balance individually because $\operatorname{tr}\Lambda = 1$ and $\lambda(w + \pi) = \lambda^*(w)$.

For $k^2 < 0$, the incoming free wave flux is associated with a wave function that is exponentially increasing in the physical region. Here ψ must be chosen so that the total free wave flux is zero, and is therefore determined by the transcendental equation

$$\int_0^\alpha (2 - e^{-i\psi}\lambda^{1/2} - e^{i\psi}\lambda^{-1/2})d\theta = 0.$$

This choice does not affect the bound pair fluxes. The state function is affected by the necessity of the inclusion of a virtual state, neither incoming nor outgoing, which exists only for that value of ψ for which the above condition holds.

As k^2 decreases past the first bound pair threshold, the flux of that bound pair must be included in the virtual state, and the condition for ψ becomes

$$8\pi^2 \cos\tau\lambda_r\lambda_r^*(\tau) + 4\pi\int_0^\alpha (2 - e^{-i\psi}\lambda^{1/2} - e^{i\psi}\lambda^{-1/2})d\theta = 0,$$

where we have chosen the first threshold to be along $\theta = 0$ with strength constant appropriate to the value of τ, which is now real.

Finally, when all thresholds have been passed the condition for a bound state is that all of the state function be in the virtual state, and ψ must satisfy

$$8\pi^2\left[\sum \cos \tau_j\lambda_r\lambda_r{}^*(\tau_j) + \sum \cos \tau_j\lambda_r\lambda_r(\alpha - \tau_j)\right] + 4\pi\int_0^\alpha (2 - e^{-i\psi}\lambda^{1/2} - e^{i\psi}\lambda^{-1/2})d\theta = 0.$$

VIII. CONCLUSION

At this point we terminate this discussion of the asymptotic solution and the algebraic formulation of three particles interacting in one dimension. Much that is said here requires further explanation. Much that is known has not been demonstrated. We have not showed, for example, how to calculate scattering and rearrangement fluxes for all values of k^2. This is a difficult calculation in general, and we have not included it because we feel that it is easier to understand in the context of a special case.

Special cases will be given in subsequent work. The principal limitation on these examples is the state of the art of algebraic computation. We hope that we have given sufficient information for the reader to see what algebra must be done.

For the asymptotic problem the fundamental algebraic function is $\lambda^{1/2}$ and its first integrals. These integrals are irreducible and specific to the algebraic structure of the problem being analyzed. One feature of this algebraic structure is the genus of the Riemann surface upon which the algebraic function may be uniformized. If $\lambda^{1/2}$ is an algebraic function that can be uniformized on a surface of genus p, there are $3p - 3$ first integrals to be computed, and these integrals cannot be expressed in terms of other more elementary functions.

The simplest class of delta function interaction problems that give rise to algebraic functions which are not elliptic functions is the case where the three masses are equal and the three strengths of interaction are nonzero, with one strength different from the other two. This problem may be reduced to 3×3 matrices, and the characteristic equation of N_μ has been worked out. It remains to calculate the appropriate integrals and to solve for the energy of the bound state and provide calculations of rearrangement and breakup probabilities. On this problem, at least, we can promise more in the future.

Another area of investigation only partially developed here is the question of the algebraic structure of the connection to the Bethe ansatz. It turns out that the necessary and sufficient condition for the consistency of the Bethe ansatz is that N_μ = identity, so that all solutions are periodic 2π in w. The algebraic condition that this occur is a set of relations among matrix elements, and the equations thus generated are called star–triangle relations. It is possible to describe particle problems that give the same star–triangle relations as those of conventional lattice problems, which establishes an algebraic relation between particle and lattice problems.

At present the only way to move from a three-particle problem to an n-particle problem is to satisfy the Bethe ansatz. The question of a generalization of the Sommerfeld ansatz to more than three particles is open.

It is a principal conclusion of this work that the Sommerfeld ansatz is consistent, and the eigenvector particular solutions exist and form a complete set for three-particle problems where the boundary conditions are not asymptotic; for periodic boundary conditions, for example. The difficulty is that the spectrum must be calculated. The difficulty in finding the law that generates the spectrum is already present in those cases where the Bethe ansatz is satisfied. When the Bethe ansatz is satisfied few spectral laws have been fully analyzed. No law for the generation of the spectrum where the Bethe ansatz fails has ever been formulated.

[1] J. B. McGuire, "Study of exactly soluble one-dimensional N-body problems," J. Math. Phys. **5**, 622 (1964).

[2] M. Gaudin, *La Fonction d'Onde de Bethe* (Masson, Paris, 1983).

[3] R. J. Baxter, *Exactly Solved Models in Statistical Mechanics*, A. P. (1982).

[4] J. B. McGuire and C. A. Hurst, "Scattering of three impenetrable particles," J. Math Phys. **13**, 1595 (1972).

[5] M. Gaudin and B. Derrida, "Solution exacte d'un probléme á trois corps. etat lié," J. Phys. **36**, 1183 (1975); B. Derrida, thesis, University of Paris, 1976.

[6] A. Sommerfeld, *Optics*, A. P. (1964), *Partial Differential Equations in Physics*, A. P. (1949); W. E. Williams, Proc. R. Soc. London Ser. A **252**, 376 (1959); H. M. Nussenzvieg, Proc. R. Soc. London Ser. A **264**, 408 (1961).

[7] L. R. Ford, *Automorphic Functions* (Chelsea, New York, 1957).

[8] F. Klein, *On Riemann's Theory of Algebraic Functions and Their Integrals* (Dover, New York, 1963).

Solitonic solutions in the Kaluza–Klein–Jordan formalism as cosmological models in general relativity

Mario C. Díaz,[a] Reinaldo J. Gleiser,[b] and Jorge A. Pullin
Facultad de Matemática Astronomía y Física Universidad Nacional de Córdoba, Laprida 854, Córdoba 5000, Argentina

(Received 12 June 1987; accepted for publication 2 September 1987)

A new renormalization procedure for the solutions of Einstein's field equations in d dimensions obtainable by the inverse scattering method (ISM) of Belinskii and Zakharov [Sov. Phys. JETP **48**, 985 (1978)] is presented. It allows one to obtain families of diagonal metrics which, in addition to the solitonic parameters that characterize the ISM, depend on $2(d-3)$ extra arbitrary parameters. An example of cosmological relevance where the source is a massless scalar field in five dimensions is presented. It represents a general family of two-parameter four-dimensional metrics that for great times go into the perfect fluid Friedman–Robertson–Walker regime.

I. INTRODUCTION

A fair amount of work has been devoted in the last years to the construction and analysis of the properties of solutions of Einstein equations with an Abelian G_2^1 group of isometries. This appears to be justified since this class contains some of the more interesting, from a physical point of view, of the stationary axially symmetric metrics as well as most of the relevant cosmological solutions. As a result, a variety of methods, such as the Bäcklund transformations, Honselaers–Kinnersley–Xanthopoulos transformations, the Neugebauer–Kramer involution, the Hauser–Ernst formulation of the Riemann–Hilbert problem, etc. (for a review see Ref. 1), have been developed and applied to the construction of solutions, starting from seed metrics with these symmetries. Among these approaches, and because of its relative simplicity, the inverse scattering method (ISM) of Belinskii and Zakharov[2] has been shown to be of great utility in the construction of models of physical relevance including superpositions of N-Kerr particles,[3] astrophysical models representing black holes distorted by surrounding matter,[4] as well as in the study of vacuum cosmological models,[5] where the ISM provides a natural way of breaking in one spatial direction the Bianchi symmetries for homogeneous models.[6,7] Interesting results have also been obtained by Carr and Verdaguer[8] and Ibañez and Verdaguer[9] concerning the propagation of gravitational waves on a Kasner background. The last two authors have also studied this type of phenomenon for a vacuum Friedmann–Robertson–Walker (FRW) model, the Milne universe.[10]

Belinskii[11] extended the ISM to study cosmological models, particularly FRW models with "stiff" fluids (i.e., a fluid where the energy density equals its pressure) and Kitchingham[12] has shown that it is possible to unify the treatment of all the models admitting two spacelike commuting Killing vectors containing massless fields (or "stiff fluids"), by casting them in the generating technique form.

A serious limitation on the class of space-times to which all these methods are applicable rests, however, on the condition that the Ricci tensor must vanish on the subspace spanned by the Killing vectors. In particular, if we assume that the energy momentum tensor corresponds to a perfect fluid, the above-mentioned restriction requires the equation of state to be that of a stiff fluid.[1] If we look, instead, for an energy momentum tensor representing an electromagnetic field then this must be of null type (i.e., such that the invariant $F^*_{ab}F^{*ab}=0$, where F^*_{ab} is the self-dual electromagnetic field tensor[13]).[14]

Progress in the direction of lifting the above-mentioned restriction started with the work of Belinskii and Ruffini[15] who extended the ISM to five-dimensional vacuum stationary axially symmetric solutions. They pointed out that through the Kaluza–Klein dimensional reduction procedure, their results were relevant to the construction of exact four-dimensional solutions with nonvanishing energy-momentum tensor.

Later Ibañez and Verdaguer[16] used these ideas to construct (starting also with a five-dimensional vacuum seed metric), cosmological solutions that represent solitonic perturbations of a FRW four-dimensional background with an effective ultrarelativistic equation of state for the matter content.

Motivated by the five-dimensional representation of the Brans–Dicke–Jordan theory of gravitation, Bruckman[17] extended the ISM to d-dimensional stationary axially symmetric space-times. He also studied families of five-dimensional solutions that reduce in the absence of rotation to the Weyl–Levi-Civita axially symmetric static vacuum metric, showing that, in certain cases, these are equivalent to the spherically symmetric solutions of the Brans–Dicke theory.

Recently,[18] we showed that, by considering five-dimensional space-times with a massless scalar field as a source, it is possible to extend the ISM to a class of perfect fluid cosmological solutions with certain symmetries. As an example we obtained finite perturbations of the solitonic type for FRW flat space-times with a perfect fluid satisfying a general "gamma" law for the equation of state.

The purpose of this paper is twofold. First we show that the ISM can be used to obtain a wider family of "solitonic" solutions than what is implied in Ref. 2. This result is based

[a] Present address: Department of Physics and Astronomy, University of Pittsburgh, Pittsburgh, Pennsylvania 15260.

[b] Member of Consejo Nacional de Investigacionnes Científicas y Técnicas of Argentina (CONICET).

0022-2488/88/010169-06$02.50

on the fact that the method does not provide directly a solution of the gravitational equations and, in general, a "renormalization" of the new (two-dimensional) metric is required in order to satisfy Einstein's equations. This problem was solved in Ref. 2 introducing a scalar "normalization" factor. We give here a generalization of this prescription that makes use of a matrix normalization factor. Second, we show that, as a result of this prescription, we obtain—when the method is applied to d-dimensional seed metrics—families of solitonic solutions that depend, in general, on $2(d-3)$ parameters (in addition to the parameters inherent to the definition of the poles of the solitonic method). They also have the interesting property that the solitonic perturbation can be made arbitrarily small.

The paper is organized as follows. We start with a review of the ISM adapted to d-dimensional nonstationary metrics with a massless scalar field acting as a source. Next, we consider the normalization problem and a possible alternative to the solution given in Ref. 2. Finally we present an example of cosmological relevance where the advantages of our entire formulation are explicitly discussed.

II. THE ISM FOR d-DIMENSIONAL NONSTATIONARY SPACE-TIMES WITH A MASSLESS SCALAR FIELD

Here we review briefly the ISM starting with a d-dimensional metric in coordinates adapted to cosmological or plane waves solutions. Because of the $d-2$ group of symmetries present, this can be written in the form

$$ds^z = e^{\Lambda(t,r)}(dr^2 - dt^2) + \mathsf{G}_{AB}(t,r)dx^A dx^B, \tag{2.1}$$

where $A,B = 1,\dots,d-2$.

In this case, generalizing the results of Tabensky and Taub[19] and Wainwright *et al.*[20] the Einstein equations for a massless scalar field as source are

$$\mathsf{R}_{\mu\nu} = \chi_{,\mu}\chi_{,\nu}, \quad \mu,\nu = 1,\dots,d. \tag{2.2}$$

If we choose null coordinates,

$$t = \xi - \eta, \quad r = \xi + \eta, \tag{2.3}$$

and write the function Λ in the form

$$\Lambda = \Lambda_V + \Lambda_M,$$

where e^{Λ_V} corresponds to a vacuum solution, we have the following splitting of Eqs. (2):

$$(\alpha\mathsf{G}_{,\xi}\mathsf{G}^{-1})_{,\eta} + (\alpha\mathsf{G}_{,\eta}\mathsf{G}^{-1})_{,\xi} = 0, \tag{2.4a}$$

$$\Lambda_{V,\xi} = \frac{(\ln\alpha)_{,\xi\xi}}{(\ln\alpha)_{,\xi}} + \frac{\mathrm{Tr}\,\mathbb{A}^2}{4\alpha\alpha_{,\xi}},$$

$$\Lambda_{V'\eta} = \frac{(\ln\alpha)_{,\eta\eta}}{(\ln\alpha)_{,\eta}} + \frac{\mathrm{Tr}\,\mathbb{B}^2}{4\alpha\alpha_{,\eta}}, \tag{2.4b}$$

where $\alpha^2 = \det \mathsf{G}_{AB}$ and the matrices $\mathbb{A}$ and $\mathbb{B}$ are

$$\mathbb{A} = -\alpha\mathsf{G}_{\xi}\mathsf{G}^{-1}, \quad \mathbb{B} = -\alpha\mathsf{G}_{,\eta}\mathsf{G}^{-1}.$$

The hydrodynamic equation for χ is

$$(\alpha\chi_{,\xi})_{,\eta} + (\alpha\chi_{,\eta})_{,\xi} = 0, \tag{2.5}$$

and the "matter function" Λ_M satisfies

$$\Lambda_{M,\xi} = (\chi_{,\xi})^2/(\ln\alpha)_{,\xi}, \quad \Lambda_{M,\eta} = (\chi_{,\eta})^2/(\ln\alpha)_{,\xi}. \tag{2.6}$$

The Lax pair for Eq. (2.4a) is

$$\mathsf{D}_1\Psi = \mathbb{A}\Psi/(\lambda-\alpha), \quad \mathsf{D}_2 = \mathbb{B}\,\Psi/(\lambda+\alpha), \tag{2.7a}$$

where $\Psi = \Psi(\xi,\eta,\lambda)$ and

$$\begin{aligned}\mathsf{D}_1 &= \partial_\xi - [2\alpha_{,\xi}\lambda/(\lambda-\alpha)]\partial_\lambda,\\ \mathsf{D}_2 &= \partial_\eta + [2\alpha_{,\eta}\lambda/(\lambda+\alpha)]\partial_\lambda.\end{aligned} \tag{2.7b}$$

Here Ψ is a $(d-2)\times(d-2)$ matrix and λ is a complex spectral parameter. Also, Ψ satisfies the boundary condition $\Psi(\lambda=0) = \mathsf{G}$.

From the Ψ_0 matrix associated with a given seed metric G_0 one then constructs the vectors

$$m_a^{(k)} = (m_0)_c{}^{(k)}[\Psi_0{}^{-1}(\mu_k,t,r)]_{ca}, \tag{2.8}$$

where $(m_0)_c{}^{(k)}$ are arbitrary real or complex parameters $(k = 1,\dots,n)$ and we assume we are interested in the construction of n-soliton solutions.

The equations for the pole trajectories that characterize the solitonic behavior of the metric are

$$\mu_{k,\xi} = 2\alpha_{,\xi}\mu_k/(\alpha-\mu_k), \quad \mu_{k,\eta} = 2\alpha_{,\eta}\mu_k/(\alpha+\mu_k), \tag{2.9}$$

where $k = 1,\dots,n$ and n is the number of solitons.

If one restricts to real pole trajectories, then the parameters also have to be real. The next step is the construction of the $n\times n$ matrix

$$\Gamma_{kl} = m_c{}^{(k)}(\mathsf{G}_0)_{cb}m_b{}^{(l)}/\mu_k\mu_l - \alpha^2, \tag{2.10}$$

and its inverse D_{kl}, such that

$$\Gamma_{kl}D_{lj} = \delta_{kj}. \tag{2.11}$$

The equation[2]

$$\mathsf{G}_{AB}{}^N = (\mathsf{G}_0)_{AB} - \sum_{kl}\frac{D_{kl}m_c{}^{(k)}(\mathsf{G}_0)_{CA}m_D{}^{(l)}(\mathsf{G}_0)_{DB}}{\mu_k\mu_l} \tag{2.12}$$

provides then a new solution of (4a). Notice, however, that this is not necessarily a solution of Einstein's equations because, in general, $\det \mathsf{G}^N{}_{CB} = \beta^2$ is not equal to α^2.

III. THE RENORMALIZATION OF THE MATRIX G^N

The problem of the normalization of G^N was solved in the general case in Ref. 2 by multiplying by an appropriate scalar factor to obtain the correct determinant for the metric.

Belinskii and Zakharov observed that taking the trace of (2.4a) it can be seen that $\det \mathsf{G}^N$ satisfies the equation

$$[\alpha(\ln\det\mathsf{G}^N)_{,\xi}]_{,\eta} + [\alpha(\ln\det\mathsf{G}^N)_{,\eta}]_{,\xi} = 0. \tag{3.1}$$

Then, it is easily verified that the matrix

$$\mathsf{G}^{ph} = [\alpha^{-2}(\det\mathsf{G}^N)]^{-1/(d-2)}\mathsf{G}^N \tag{3.2}$$

satisfies both (2.4a) and the condition $\det \mathsf{G}^{ph} = \alpha^2$.

The same ansatz was used by Verdaguer,[7] and later by Bruckman,[17] for the d-dimensional metrics. However, in the diagonal case, it is possible to make another choice, which allows one to obtain a more general family of solutions. The main idea is to multiply the matrix G^N by a $(d-2)\times(d-2)$ diagonal matrix N so that

$$\mathsf{G}_{AB}{}^{ph} = \mathsf{N}_{AC}\mathsf{G}^N{}_{CB} \tag{3.3}$$

becomes a solution of Einstein's equations with $\det \mathsf{G}^{ph} = \alpha^2$.

Because G^N is a solution of (2.4a) we have to look for a matrix N that satisfies

$$(\alpha \mathsf{N}_{,\xi}\mathsf{N}^{-1})_{,\eta} + (\alpha \mathsf{N}_{,\eta}\mathsf{N}^{-1})_{,\xi} = 0. \tag{3.4}$$

The following structure for the matrix N guarantees that this equation will be satisfied:

$$\begin{aligned}&\mathsf{N}_{ii} = \alpha^{p_i}\beta^{q_i}, \quad i = 1,...,d-3,\\ &\mathsf{N}_{d-2,d-2} = \alpha^{2-P}\beta^{-2-Q}, \quad P = \sum_{i=1}^{d-3} p_i, \quad Q = \sum_{i=1}^{d-3} q_i,\\ &\mathsf{N}_{kj} = 0, \quad k \neq j.\end{aligned} \tag{3.5}$$

The reason is that $\beta = \det \mathsf{G}^N$ satisfies (3.1) and α is a solution of the wave equation ($\alpha_{,\xi\eta} = 0$).

Notice that with the choice $p_i = -q_i = 2/(d-2)$, we recover the standard prescription.[2,7,17] It should also be mentioned that although this nonuniqueness of the normalization required to go from G^N to G^{ph} does not seem to have been analyzed before, Letelier[21] arrived at similar results (for $d = 4$) by direct integration of (2.4a) in the diagonal static case. We remark, however, that what our construction indicates is that we should consider as solitonic a somewhat wider range of solutions than that implied in Ref. 2.

An interesting question is the possible generalization of these results to nondiagonal metrics. We notice that some restrictions must be placed on N, for otherwise the problem would, in a way, become trivial. In particular, we would like to recover, in the general case, the possibility discussed below of making the solitonic perturbation arbitrarily small. This problem is currently under study.

To obtain the full solution we need now to integrate (2.4b). It is a remarkable fact that even with the normalization (3.5), the function $\exp(\Lambda_V)$ can be obtained directly by quadratures from the matrix $\mathsf{G}^{ph}{}_{AB}$. (Some relations between the functions μ_k and their derivatives that allow to integrate the solution easily are listed in the Appendix.)

In general the n-soliton solution for matrix G will have the form

$$\mathsf{G}_{ii} = t^{2\delta_i}\left(\prod_{k=1}^{n} \mu_k\right)^{\gamma_i} r^{2\beta_i}(rt)^{2\epsilon_2^i}, \quad i = 1,...,d-2, \tag{3.6}$$

where δ_i, γ_i, and β_i are real exponents restricted only by the requirements

$$\sum_i \delta_i = 0, \quad \sum_i \beta_i = 0, \quad \sum_i \gamma_i = 0, \tag{3.7}$$

which are automatically satisfied by the procedure we explained above, and ϵ_j^i is the Kronecker tensor.

The relations

$$\begin{aligned}&t^2 = T - (T^2 - R^2)^{1/2} \equiv -\mu_T,\\ &r^2 = T + (T^2 - R^2)^{1/2} \equiv -\mu_R,\end{aligned} \tag{3.8}$$

can be used to define new coordinates R, T. From (2.4b), and using the equations given in the Appendix, we have, in these coordinates,

$$\begin{aligned}e^{\Lambda_V} = {}& C(\mu_T)^{\delta_2}(\mu_R)^{\beta_2}\left(\prod_{k=1}^{n} \mu_k\right)^{\gamma_2}(\mu_T - \mu_R)^{\Sigma\delta_i\beta_i}\left[\prod_{k=1}^{n} (\mu_T - \mu_k)\right]^{\Sigma\delta_i\gamma_i}\left[\prod_{k=1}^{n} (\mu_k - \mu_r)\right]^{\Sigma\beta_i\gamma_i}\left[\prod_{k>j}^{n} (\mu_k - \mu_j)\right]^{\Sigma\gamma_i^2}\\ &\times\left[\prod_{k=1}^{n} \frac{\mu_k{}^2}{R^2 - \mu_k{}^2}\right]^{\Sigma_i(\gamma_i^2/2)}\left[\frac{\mu_T{}^2}{R^2 - \mu_T{}^2}\right]^{\Sigma_i(\delta_i^2/2)}\left[\frac{\mu_R{}^2}{R^2 - \mu_R{}^2}\right]^{\Sigma_i(\beta_i^2/2)}.\end{aligned} \tag{3.9}$$

Now using (3.8) we may write the corresponding coefficient in the r,t coordinates. The result is

$$e^{\Lambda_V} = Ct^L r^M (r^2 - t^2)^N \left(\prod_{k=1}^{n} \mu_k\right)^{\gamma_2}\left[\prod_{k=1}^{n} (t^2 + \mu_k)\right]^{\Sigma\delta_i\gamma_i}\left[\prod_{k=1}^{n} (r^2 + \mu_k)\right]^{\Sigma\beta_i\gamma_i}\left[\prod_{k>j}^{n} (\mu_k - \mu_j)\right]^{\Sigma\gamma_i^2}\left[\prod_{k=1}^{n} \frac{\mu_k{}^2}{t^2r^2 - \mu_k{}^2}\right]^{\Sigma_i(\gamma_i^2/2)}, \tag{3.10}$$

where

$$\begin{aligned}&L = \sum_{i=1}^{d-2} \delta_i{}^2 + 2\delta_2, \quad M = \sum_{i=1}^{d-2} \beta_i{}^2 + 2\beta_2,\\ &N = 1 - \frac{1}{2}\sum_{i=1}^{d-2} (\delta_i - \beta_i)^2.\end{aligned}$$

The potential χ remains unperturbed by the ISM.[11] Therefore we only have to multiply (3.10) by the factor $\exp(\Lambda_M)$ determined by Eqs. (2.6) to obtain the new function $e^{\Lambda(t,r)}$.

These solutions provide a $2(d-3)$-parameter family of metrics of the d-dimensional Einstein–Rosen type, satisfying the d-dimensional Einstein equations (2.2) for a massless scalar field. In accordance with the previous discussion, the metrics are of solitonic type. What we are interested in is, however, the construction of four-dimensional solutions of Einstein's equations with nonvanishing energy-momentum tensor. In the present context this can be achieved through the Kaluza–Klein–Jordan procedure of reference.[15] The following example will be useful to illustrate the solitonic character of the solutions as well as the role played by the massless scalar field that appears in the d-dimensional energy momentum tensor.

IV. FINITE SOLITONIC PERTURBATIONS OF FRW MODELS

We take as seed metric a flat FRW space-time corresponding to a perfect fluid whose equation of state obeys a general γ law

$$p=(\gamma-1)\varepsilon,$$
$$ds^2=t^{n}(-dt^2+dr^2+r^2d\varphi^2+dz^2), \tag{4.1}$$

where n is a real constant, r, φ, and z are cylindrical coordinates, and t is the conformal time. Here γ and n are related by

$$\gamma=2(2+n)/(3n).$$

From (4.1) we obtain a related five-dimensional metric of the form (2.1) defining $\mathrm{G}_{33}=\phi^2$ and $\mathrm{G}_{A3}=0$ ($A=1,2$). If we look for solutions of the form (2.2) for the five-dimensional Ricci tensor we have an effective four-dimensional energy-momentum tensor of the form

$$T_{\mu\nu}=\phi^{-1}\phi_{;\mu;\nu}+\chi_{,\mu}\chi_{,\nu}-\tfrac{1}{2}g_{\mu\nu}(\chi^{\alpha}\chi_{\alpha}). \tag{4.2}$$

It is an interesting fact[16] that for the flat FRW space-time the function ϕ can be interpreted as the potential for a radiative fluid with an energy density defined by the equation $\varepsilon_R(4\mu_\mu\mu_\nu+g_{\mu\nu})=3\phi^{-1}\phi_{;\mu;\nu}$, where u^μ is the four-velocity of the fluid. The function χ, on the other hand, may be thought of as the potential for an irrotational stiff fluid with energy density $\varepsilon_S=-\tfrac{1}{2}\chi_{,\alpha}\chi_{,\beta}g^{\alpha\beta}$ (Refs. 11 and 19).

Besides, as was pointed in Ref. 18, it is always possible to make a (formal) unique decomposition of the energy-momentum tensor of any perfect fluid of the form

$$T_{\mu\nu}=(\varepsilon+p)\mu_\mu\mu_\nu+pg_{\mu\nu}$$

into a sum of (perhaps unphysical) radiative and stiff fluids. The importance of this fact is that, when a potential ϕ for the radiative fluid can be found, it leads to a modification of the formalism used in Ref. 16 which makes the ISM applicable (in the context of a five-dimensional Kaluza–Klein–Jordan formalism) to a class of perfect fluid solutions of Einstein's equations. In what follows we refer the reader to Ref. 18 for details. The relevant quantities that define the seed metric in the present example are

$$\phi=\mathrm{G}_{33}{}^{1/2}=t^{-n+1},\quad \alpha^2=t^2r^2,\quad \varepsilon=\tfrac{3}{4}n^2t^{-n-2},$$
$$\chi=[\tfrac{3}{2}n(2-n)]^{1/2}\ln t,\quad e^{\Lambda_M}=\left[\frac{t^2}{r^2-t^2}\right]^{3n(2-n)/4}. \tag{4.3}$$

From this metric, with the appropriate choice of the constant vectors $m_{0c}{}^{(k)}$, we can obtain the following three-dimensional reduced two-solution metric:

$$\mathrm{G}^N_{AB}=\begin{bmatrix} t^{n} & & \\ & t^{n}r^2 & \\ & & t^{-2n+2}\alpha^4/\mu_1^2\mu_2^2\end{bmatrix}, \tag{4.4}$$

where the functions μ_k are

$$\mu_k{}^{\pm}=-\omega_k{}^2-t^2-r^2\pm\{(\omega_k^2+r^2+t^2)^2-4t^2r^2\}^{1/2}. \tag{4.5}$$

In this case we have $\beta=t^{\sigma}r^{\sigma}/(\mu_1\mu_2)^2$ and then a physical matrix for the problem is

$$\mathrm{G}^{ph}=\begin{bmatrix} \dfrac{t^{n+p_1+q_1}r^{p_1+q_1}}{(\sigma_1\sigma_2)^{q_1}} & & \\ & \dfrac{t^{n+p_2+q_2}r^{2+p_2+q_2}}{(\sigma_1\sigma_2)^{q_2}} & \\ & & \dfrac{t^{-2n+2-p}r^{-p}}{(\sigma_1\sigma_2)^{-(q_1+q_2)}}\end{bmatrix}, \tag{4.6}$$

where $p=p_1+q_1+p_2+q_2$, and $\sigma_1=\sigma_1^-$, $\sigma_2=\sigma_2^+$ with $\sigma_k^{\pm}=\mu_k^{\pm}/(2rt)$.

For $t\to\infty$ these functions have the behavior,

$$\sigma_k^+\underset{t\to\infty}{\cong}-r/t+\omega_k{}^2r/t^3,\quad \sigma_k^-\underset{t\to\infty}{\cong}-t/r+\omega_k{}^2/rt. \tag{4.7}$$

Therefore, for $r\ll t\to\infty$, we have

$$\sigma_1\sigma_2\underset{r\ll t\to\infty}{\cong}1+(\omega_1{}^2-\omega_2{}^2)/t^2,$$

and in this case G^{ph} goes to the background if we choose $p_i=q_i$. With this election of parameters we have

$$\mathrm{G}^{ph}=\begin{bmatrix} t^{n}/(\sigma_1\sigma_2)^{q_1} & & \\ & t^{n}r^2/(\sigma_1\sigma_2)^{q_2} & \\ & & t^{-2n+2}(\sigma_1\sigma_2)^{q_1+q_2}\end{bmatrix}, \tag{4.8}$$

and, using (17),

$$e^{\Lambda}=\frac{Ct^{n-2}(\sigma_1-\sigma_2)^{\Sigma\gamma_i^2-2}(\sigma_1\sigma_2)^{-q_2+2}[(t+\sigma_1r)(t+\sigma_2r)]^{Q}}{r^2[(1-\sigma_1{}^2)(1-\sigma_2{}^2)]^{\Sigma\gamma_i^2/2}(1-\sigma_1\sigma_2)^2}, \tag{4.9}$$

where $i=1,2,3$, and

$$\delta_1=n/2+q_1,\quad \beta_1=-\gamma_1=q_1,\quad \delta_2=n/2+q_2-1,\quad \beta_2=-\gamma_2=q_2,$$
$$\delta_3=-(n-1+q_1+q_2),\quad \beta_3=-(q_1+q_2)=-\gamma_3,\quad Q=\sum_{i=1}^{3}\gamma_i(\delta_i-\beta_i)=-\frac{q_1(3n-2)}{2}-\frac{q_2(3n-4)}{2}.$$

It is easy to verify that for $q_1 = q_2 = -\frac{2}{3}$ we recover the metric of Ref. 18. With the same choice of q_1 and q_2 but for $n = 2$ (when the background corresponds to a radiative fluid) we have metric (9) of Ref. 16. Similarly, with $q_1 = -\frac{4}{3}$ and $q_2 = \frac{2}{3}$ we obtain metric (10) of Ref. 16.

In general we have a two-parameter family of metrics that represent cylindrical perturbations of solitonic type on the background of flat FRW universes. A relevant difference with respect to the cases of Ref. 16 and 18 is that now we can choose $q_1 = -q_2$ and deal with a new solution where the function ϕ remains unperturbed. In this case we can construct a one-parameter subfamily of metrics with G_{33} unperturbed. These are given by

$$\mathsf{G}_{ph} = \begin{bmatrix} \dfrac{t^{n}}{(\sigma_1\sigma_2)^q} & & \\ & t^{n} r^2 (\sigma_1\sigma_2)^q & \\ & & t^{-2n+2} \end{bmatrix}, \tag{4.10}$$

$$e^{\Lambda} = \frac{C t^{n-2} (\sigma_1 - \sigma_2)^{2(q^2-1)} (\sigma_1\sigma_2)^{2+q} [(t+\sigma_1 r)(t+\sigma_2 r)]^{-q}}{r^2 [(1-\sigma_1{}^2)(1-\sigma_2{}^2)]^{q^2} (1-\sigma_1\sigma_2)^2}. \tag{4.11}$$

The limit of this metric for $q \to 0$ is the seed metric as can be easily seen with the aid of Eq. (A6). Thus, since near $q = 0$ the metric is a smooth function of the parameter q, we can make the solitonic perturbation as small as we want. Actually, this interesting feature is a consequence only of our normalization procedure. Clearly, it may be helpful, e.g., in the analysis of the stability of the seed metric under small perturbations of solitonic type.

Regarding the general solution, we may point out that, as in Ref. 18, we have the following interesting cases for the value of n.

The $n = 1$ case corresponds to a stiff fluid background similar to that studied by Belinskii.[11]

For $n = 2$ we deal with the radiative case studied by Ibañez and Verdaguer.[16]

For $n = -2$ we have a vacuum de Sitter background, the behavior of which is studied in Ref. 22 for the particular values of q_i, $q_1 = q_2 = -\frac{2}{3}$. This case is particularly interesting in relation to the study of the stability of inflationary models.

For $n = 0$ the background is the Minkowski space time.

For $n = 4$ the background is dust or pressureless perfect fluid.

We remark that in all these cases not only is the metric but also the matter content perturbed and, in general, we do not have as source an energy-momentum tensor corresponding to a perfect fluid. This is the case even when the potential $\phi = [\mathsf{G}_{33}]^{1/2}$ remains unperturbed, because the energy momentum tensor is defined via the covariant derivatives of the potential, which are computed with the perturbed metric.

In general, the energy-momentum tensor describes an anisotropic fluid that can be made to satisfy the strong and dominant energy conditions,[23] at least in some region of the space-time, provided the appropriate election of the relation between ω_1 and ω_2 is made. Another gauge freedom, namely the election of the constant C, can be fixed by the requirement of regularity of the metric on the symmetry axis.[13]

V. CONCLUSIONS

We have shown that it is possible to generalize the renormalization procedure for the solutions obtainable with the ISM. In spite of its simplicity, this generalization leads to a richer class of solitonic perturbations of the given seed metric than the standard prescription because, for every possibility of election of the constant vectors $(m_0)^{(k)}_c$ in (2.8), we obtain now, in general, a $2(d-3)$-parameter family of metrics. This feature may be helpful in the study of perturbations of cosmological models. The reason is that, although the symmetry is rather restrictive, these perturbations can be made of arbitrary size and they are due to their construction of nonlinear type, therefore closely resembling the processes that may occur in nature.

Another aspect of our presentation is the use of the ISM for the construction of d-dimensional space-times with a massless scalar field as source for the matter content. These have shown to be of great utility, at least in one case—the flat FRW model—where with its aid it was possible to construct solitonic perturbations of perfect fluid solutions in four dimensions. This result provides interest and justification to the search of other fruitful ideas leading to its application to a wider class of space-times with matter.

A more detailed analysis of the properties of the family of metrics we presented in Sec. IV is under consideration.[24]

ACKNOWLEDGMENTS

M. C. D. and J. A. P. deeply acknowledge the hospitality and partial support from Instituto Balseiro, Centro Atomico Bariloche, where part of this work was done, especially the hospitality of Professor Victor Hugo Ponce and Professor Oscar Bressan. J. A. P. also acknowledges the support from the Consejo Nacional de Investigaciones Científicas y Técnicas of Argentina (CONICET) through a fellowship.

APPENDIX

The following relations may be useful in computing the metric coefficients:

$$\left[\ln\left(\frac{\mu_i{}^2}{r^2 - \mu_i{}^2}\right)\right]_{,r} = \frac{r}{2}\left[\left(\frac{\mu_{i,r}}{\mu_i}\right)^2 + \left(\frac{\mu_{i,t}}{\mu_i}\right)^2\right], \tag{A1}$$

$$\left[\ln\left(\frac{\mu_i{}^2}{r^2 - \mu_i{}^2}\right)\right]_{,t} = r\frac{\mu_{i,r}}{\mu_i}\frac{\mu_{i,t}}{\mu_i}, \tag{A2}$$

$$[\ln(\mu_i - \mu_j)^2]_{,r} = r\,\frac{\mu_{i,t}\mu_{j,t} + \mu_{i,r}\mu_{j,r}}{\mu_i\mu_j}, \tag{A3}$$

$$[\ln(\mu_i - \mu_j)^2]_{,t} = r\frac{\mu_{i,r}\mu_{j,t} + \mu_{i,t}\mu_{j,r}}{\mu_i\mu_j}. \tag{A4}$$

We also have

$$\sigma_1\sigma_2/[(t+\sigma_1 r)(t+\sigma_2 r)(r+\sigma_1 t)(r+\sigma_2 t)] = a, \tag{A5}$$

$$(\sigma_1 - \sigma_2)(1-\sigma_1\sigma_2)tr/\sigma_1\sigma_2 = b, \tag{A6}$$

where ab and ω_k are real constants, and

$$\mu_k = \omega_k - t \pm \{(\omega_k - t)^2 - r^2\}^{1/2}.$$

[1] M. A. H. MacCallum, "Exact solutions in cosmology," in *Solutions of Einstein's Equations: Techniques and Results, Lecture Notes in Physics*, Vol. 205, edited by C. Hoenselaers and W. Dietz (Springer, Berlin, 1984), p. 334.

[2] V. A. Belinskii and V. E. Zakharov, Sov. Phys. JETP **48**, 985 (1978).

[3] V. A. Belinskii and V. E. Zakharov, Sov. Phys. JETP **50**, 1 (1979).

[4] A. Tomimatsu, Phys. Lett. A **103**, 374 (1984).

[5] V. A. Belinskii and M. Francaviglia, Gen. Relativ. Gravit. **14**, 213 (1982).

[6] R. T. Jantzen, Nuovo Cimento B **59**, 287 (1980).

[7] E. Verdaguer, *Observational and Theoretical Aspects of Relativistic Astrophysics and Cosmology*, edited by J. L. Sanz and L. J. Goicochea (World Scientific, Singapore, 1985).

[8] B. J. Carr and E. Vergaguer, Phys. Rev. D **28**, 2995 (1983).

[9] J. Ibáñez and E. Verdaguer, Phys. Rev. D **31**, 251 (1985).

[10] J. Ibáñez and E. Verdaguer, Class. Quantum Gravit. **3**, 1235 (1986).

[11] V. A. Belinskii, Sov. Phys. JETP **50**, 623 (1979).

[12] D. W. Kitchingham, Class. Quantum Gravit. **3**, 133 (1986).

[13] D. Kramer, H. Stephani, M. MacCallum, and E. Herlt, *Exact Solutions of Einstein's Field Equations* (Cambridge U.P., Cambridge, 1980).

[14] V. A. Belinskii, JETP Lett. **30**, 29 (1979).

[15] V. A. Belinskii and Ruffini, Phys. Lett. B **89**, 197 (1980).

[16] J. Ibáñez and E. Verdaguer, Astrophys. J. **306**, 401 (1986).

[17] W. Bruckman, Phys. Rev. D **34**, 2990 (1986).

[18] M. C. Díaz, R. J. Gleiser, and J. A. Pullin, Class. Quantum Gravit. **4**, L23 (1987).

[19] R. Tabensky and A. H. Taub, Commun. Math. Phys. **29**, 61 (1973).

[20] J. Wainwright, W. C. W. Ince, and B. J. Marshman, Gen. Relativ. Gravit. **10**, 259 (1979).

[21] P. Letelier, J. Math Phys. **26**, 467 (1985).

[22] M. C. Díaz, R. J. Gleiser, and J. A. Pullin, submitted for publication.

[23] S. W. Hawking and G. F. R. Ellis, *The Large Scale Structure of Space-time* (Cambridge U.P., Cambridge, 1973).

[24] M. C. Díaz, Ph.D. thesis, University of Cordoba, 1987.

Type-D rigidly rotating perfect fluid solutions

Alberto García D. and Isidore Hauser[a)]

Centro de Investigación y de Estudios Avanzados del I. P. N., Departamento de Física, Apdo. Postal 14-740, 07000, México, D. F., Mexico

(Received 12 March 1987; accepted for publication 9 September 1987)

Within the type-D Carter metric structure with a perfect fluid source, it is established that the Wahlquist metric and a new divergenceless solution are the only perfect fluid solutions satisfying positive energy and regularity conditions. The divergenceless solution, obtainable also as a limiting transition of the Wahlquist one, is endowed with three arbitrary parameters, and contains the Kramer fluid solution as a particular case. The fluid four-velocity of these metrics does not lie in the two-space spanned by the double principal null directions of the Weyl tensor.

I. INTRODUCTION

The first perfect fluid solution for a rotating fluid body bounded by a finite surface of zero pressure has been obtained by Wahlquist[1] in 1968. This solution is of type D with geodesic, shear-free, twisting, and diverging principal null directions of the Weyl conformal curvature tensor. Several years later, in 1984, Kramer derived a type-D rotating perfect fluid solution with geodesic, shear-free, twisting, and divergenceless principal null directions.[2] One can easily establish that both the Wahlquist and the Kramer solutions belong to the type-D Carter metric[3]; see metric (1) of Ref. 3, which we shall call Carter metric from now on. In the Wahlquist's case, this fact is shown in Ref. 4. Thus it is reasonable to expect that the Kramer solution could be derived from the Wahlquist one by a limiting process; in fact, this happens to be the case as we shall show in this work.

The main goal of this paper is to demonstrate that within the type-D Carter metric structure there exist two physically meaningful (satisfying regularity and energy conditions) branches of solutions; the Wahlquist metric and a three-parameter divergenceless solution, which contain as a particular case the Kramer metric.

The second purpose is to show that our new divergenceless fluid solution can be obtained from the diverging Wahlquist metric by a limiting contraction process.

While this paper was being reviewed,[5] we learned of the work by Kramer[6] on closely related subjects, namely, on the determination of the divergenceless fluid solution and the interpretation of the Kramer solution as a limiting case of the Wahlquist one.

II. EQUATIONS FOR TYPE-D CARTER ROTATING PERFECT FLUID SOLUTIONS

The Carter metric occupies a remarkable place in the theory of exact solutions. It contains—modulo limiting transitions—all type-D solutions in both vacuum and electrovac cases except for solutions endowed with the acceleration parameter. It was derived by Carter in 1968 assuming that the space-time allows (i) a two-parameter Abelian invertible with non-null surfaces of transitivity symmetry group (time independence and axial symmetry), (ii) the separability of the Hamilton–Jacobi equation for geodesics, and (iii) the separability of the Schrödinger equation [when (iii) holds (ii) does automatically].

The stationary axisymmetric Carter line element, formula (1) of Ref. 3, can be given in the real chart $\{x^\mu\} = \{x,y,\sigma,\tau\}$ as

$$\begin{aligned} g = {} & (\Delta/P)dx^2 + (P/\Delta)(d\tau + N\,d\sigma)^2 \\ & + (\Delta/Q)dy^2 - (Q/\Delta)(d\tau + M\,d\sigma)^2, \\ & P = P(x), \quad Q = Q(y), \quad M = M(x), \\ & N = N(y), \quad \Delta := M - N, \end{aligned} \tag{2.1}$$

where σ and τ are ignorable coordinates corresponding to the spacelike ∂_σ and timelike ∂_τ Killing directions. Since the signature we use is $(+\ +\ +\ -)$, we have to require $P/\Delta > 0$ and $Q/\Delta > 0$.

Considering our chart as comoving coordinates, the components of the fluid four-velocity vector are

$$\begin{aligned} & u^\mu = \delta^\mu_\tau \, (\Delta/(Q-P))^{1/2}, \\ & u^\mu u_\mu = -1, \quad (Q-P)/\Delta > 0. \end{aligned} \tag{2.2}$$

The Einstein equations

$$\begin{aligned} R_{\mu\nu} - \tfrac{1}{2} g_{\mu\nu} R = - T_{\mu\nu}, \quad T_{\mu\nu} = (\epsilon + p) u_\mu u_\nu + p g_{\mu\nu}, \\ \epsilon + p \neq 0, \quad \epsilon > 0, \end{aligned} \tag{2.3}$$

for type-D solutions yield the following.

(i) The equation for $M(x)$ and $N(y)$,

$$\begin{aligned} & \partial_x (M_x/\Delta) - \partial_y (N_y/\Delta) = 0 \\ & \to M_{xx} - N_{yy} - \Delta^{-1}[(M_x)^2 + (N_y)^2] = 0. \end{aligned} \tag{2.4}$$

(ii) The equation for the structural functions $P(x)$ and $Q(y)$,

$$\begin{aligned} & \Delta(P_{xx} - Q_{yy}) - (M_{xx} + N_{yy})(P+Q) \\ & \quad - 2M_x P_x + 2M_{xx} P - 2N_y Q_y + 2N_{yy} Q = 0. \end{aligned} \tag{2.5}$$

(iii) The equation defining the pressure p,

$$p = \tfrac{1}{4}(P_{xx} + Q_{yy})\Delta^{-1}. \tag{2.6}$$

(iv) The equation for the energy density ϵ,

$$\epsilon + p = \tfrac{1}{2}(M_{xx} + N_{yy})(Q-P)\Delta^{-2} \neq 0. \tag{2.7}$$

The double principal null directions of the conformal tensor for the studied metric (2.1) are geodesic, shear-free,

[a)] This work was carried out during September 1985–June 1986 when this author was at CINVESTAV on sabbatical leave from the Illinois Institute of Technology, Chicago, Illinois. Present address: 1615 Cottonwood Dr., #18, Louisville, Colorado 80027.

0022-2488/88/010175-04$02.50

and in general possess complex expansion Z,

$$Z = \Theta + i\rho = -(1/2\Delta)(Q/2\Delta)^{1/2}[N_y + iM_x], \qquad (2.8)$$

where Θ and ρ denote the divergence and the twist, respectively.

This geometric characteristic Z allows us to classify the viable solutions according to the following schema:

(i) twisting ($M_x \neq 0$) and diverging ($N_y \neq 0$) solutions,

(ii) twisting ($M_x \neq 0$) and divergenceless ($N_y = 0$) solutions, (2.9)

(iii) twist-free ($M_x = 0$) and diverging ($N_y \neq 0$) solutions.

Sections III–V deal with the integration of the field equations (2.4) and (2.5), for each branch of the schema above, requiring additionally that the obtained solutions have to be such that the energy density should be positive, and on the limit at the rotation axis they do satisfy the regularity condition

$$\xi_{,\mu}\xi^{,\mu}/4\xi \to 1, \quad \xi = \zeta^{\mu}\zeta_{\mu}, \quad \zeta^{\mu} = \delta^{\mu}_{\sigma}. \qquad (2.10)$$

For the sake of completeness, the expressions for the tetrad complex curvature coefficients $C^{(a)}$, $a = 1,...,5$, of the conformal Weyl tensor are given. With respect to the null tetrad

$$\left.\begin{matrix} e^1 \\ e^2 \end{matrix}\right\} = \frac{1}{\sqrt{2}}\left\{\left(\frac{\Delta}{P}\right)^{1/2} dx \pm i\left(\frac{P}{\Delta}\right)^{1/2}(d\tau + N\,d\sigma)\right\},$$

$$\left.\begin{matrix} e^3 \\ e^4 \end{matrix}\right\} = \frac{1}{\sqrt{2}}\left\{\left(\frac{\Delta}{Q}\right)^{1/2} dy \pm \left(\frac{Q}{\Delta}\right)^{1/2}(d\tau + M\,d\sigma)\right\}, \qquad (2.11)$$

the only nonvanishing $C^{(a)}$ component is $C^{(3)}$,

$$\begin{aligned} 6\Delta C^{(3)} &= P_{xx} - 3\frac{M_x}{\Delta}P_x - 2P\left[\frac{M_{xx}}{\Delta} - 2\left(\frac{M_x}{\Delta}\right)^2 + \left(\frac{N_y}{\Delta}\right)^2\right] \\ &\quad + Q_{yy} + 3\frac{N_y}{\Delta}Q_y \\ &\quad + 2Q\left[\frac{N_{yy}}{\Delta} + 2\left(\frac{N_y}{\Delta}\right)^2 - \left(\frac{M_x}{\Delta}\right)^2\right] \\ &\quad + \frac{3i}{\Delta^2}[(\Delta Q_y + 2QN_y)M_x + (\Delta P_x - 2PM_x)N_y], \end{aligned} \qquad (2.12)$$

therefore, the studied metric is of type D.

III. TWISTING AND DIVERGING SOLUTIONS; THE WAHLQUIST METRIC

In case (i) of (2.9) two different sets of real structural functions $\{M, N, P, Q\}$ satisfy the field equations (2.4) and (2.5). Nevertheless, only the set of structural functions corresponding to the Wahlquist metric gives rise to a positive energy density, while the second possibility possesses a negative energy density.

Equation (2.4), differentiated with respect to x and y, yields

$$M_{xxx}/M_x = -\alpha = -N_{yyy}/N_y, \quad \alpha = \text{const.} \qquad (3.1)$$

If α is equal to zero, without loss of generality, the general solution of (3.1), fulfilling also (2.4), can be given as $M = ax^2 + b$, and $N = -ay^2 + b$, where a and b are arbitrary constants. These structural functions imply, according with (2.7), that $p + \epsilon = 0$, and therefore they give rise to solutions outside the class we are interested in.

For α different from zero, one can distinguish two branches of solutions,

$$\begin{aligned} &\text{(a)} \quad \alpha = -4\nu^2, \quad M = \kappa\cosh 2\nu x + \mu, \\ &\qquad N = \kappa\cos 2\nu y + \mu, \end{aligned} \qquad (3.2)$$

$$\begin{aligned} &\text{(b)} \quad \alpha = 4\nu^2, \quad M = \kappa\cos 2\nu x + \mu, \\ &\qquad N = \kappa\cosh 2\nu y + \mu, \end{aligned} \qquad (3.3)$$

where μ, ν, and κ are constants.

For $\alpha = -4\nu^2$, the Wahlquist case, we notice that

$$M_{xx} + N_{yy} = 4\nu^2\Delta, \qquad (3.4)$$

and hence Eq. (2.5) for P and Q becomes

$$\begin{aligned} &\Delta(P_{xx} - 4\nu^2 P) - 2M_xP_x + 2M_{xx}P \\ &\quad - \Delta(Q_{yy} + 4\nu^2 Q) - 2N_yQ_y + 2N_{yy}Q = 0, \end{aligned} \qquad (3.5)$$

which differentiated with respect to x and y yields

$$\begin{aligned} &(1/\sinh 2\nu x)\partial_x(P_{xx} - 4\nu^2 P) \\ &\quad = 8\beta\nu = (1/\sin 2\nu y)\partial_y(Q_{yy} + 4\nu^2 Q), \end{aligned} \qquad (3.6)$$

where β is a separation constant.

Integrating these equations, one arrives at

$$\begin{aligned} &P_{xx} - 4\nu^2 P = 4\beta\cosh 2\nu x - 4a\nu^2, \quad a = \text{const}, \\ &Q_{yy} + 4\nu^2 Q = -4\beta\cos 2\nu y + 4c\nu^2, \quad c = \text{const}. \end{aligned} \qquad (3.7)$$

The general solution of (3.7), fulfilling also (2.5), can be given as

$$\begin{aligned} &P = a + b\cosh 2\nu x - n\sinh 2\nu x + (\beta x/\nu)\sinh 2\nu x, \\ &Q = a + b\cos 2\nu y - m\sin 2\nu y - (\beta y/\nu)\sin 2\nu y, \end{aligned} \qquad (3.8)$$

where a, b, n, and m are integration constants.

Evaluating p and ϵ from (2.6) and (2.7) one obtains

$$\begin{aligned} &p = -(\nu^2/\Delta)(Q - P) + \beta/\kappa, \\ &\epsilon = 3(\nu^2/\Delta)(Q - P) - \beta/\kappa, \quad \epsilon + 3p = 2(\beta/\kappa). \end{aligned} \qquad (3.9)$$

Consequently, on the zero pressure surface ($p = 0$), the value of the surface energy amount to $\epsilon_s = 2(\beta/\kappa) > 0$. If $\beta = 0$, one arrives at the Vaidya metric.[7] One can show, by scaling transformations, that the obtained metric possesses only five free parameters; for their interpretation see Ref. 1.

The "formal" solution of case (b), with negative energy density, is obtainable from the Wahlquist metric replacing ν by $i\nu$, $\nu \to i\nu$.

IV. TWISTING AND DIVERGENCELESS PERFECT FLUID SOLUTION

In case (ii) of (2.9) $N_y = 0 \neq M_x$, the general integral of (2.4) is given by

$$M = \kappa e^{2\nu x} + l, \quad N = l, \quad \nu \neq 0 \neq \kappa, \qquad (4.1)$$

where κ, ν, and l are real constants.

Substituting M, N, and their derivatives into (2.5), one arrives at a variable separable equation for P and Q, namely,

$$P_{xx} - 4\nu P_x + 4\nu^2 P = 4\nu^2\mu = Q_{yy} + 4\nu^2 Q, \quad \mu = \text{const.} \tag{4.2}$$

The general integral of (4.2) can be given as

$$P = \mu + e^{2\nu x}(ax + b), \quad Q = \mu + \gamma\cos(2\nu y - \beta), \tag{4.3}$$

where a, b, β, and γ are integration constants. Without any loss of generality one can set $\beta = 0$; we shall adopt this choice of β from now on.

One easily evaluates p and ϵ, which amount to

$$\begin{aligned} p &= -(\nu^2/\Delta)(Q-P) + \nu a/\kappa\,, \\ \epsilon &= 3(\nu^2/\Delta)(Q-P) - \nu a/\kappa\,, \quad \epsilon + 3p = 2(\nu a/\kappa)\,. \end{aligned} \tag{4.4}$$

Hence the energy on the zero pressure surface amounts to $\epsilon_s = 2(\nu a/\kappa) > 0$.

The evaluation of $C^{(3)}$ from (2.12) yields

$$C^{(3)} = -(\nu/3\Delta)\{ae^{2\nu x} + 6\nu\gamma e^{2i\nu y}\}. \tag{4.5}$$

From (4.4) and (4.5) one arrives at the conclusion that the derived solution is equipped with three essential parameters; ν, a, and γ. Notice that if a equal to zero, we have a "solution" for which $\epsilon + 3p = 0$ with positive energy density but with negative pressure.

By linear transformations of the form

$$x \to \alpha x + x_0, \quad \tau \to \beta\tau + b\sigma, \quad y \to \alpha y + y_0, \quad \sigma \to c\sigma, \tag{4.6}$$

the studied metric, assuming $\epsilon + 3p \neq 0$, can be brought to the form

$$\begin{aligned} g &= \frac{\Delta}{P}\,dx^2 + \frac{P}{\Delta}\,d\tau^2 + \frac{\Delta}{Q}\,dy^2 - \frac{Q}{\Delta}(d\tau + \Delta\,d\sigma)^2, \\ \Delta &= e^x, \quad P = \mu + \delta(x + \epsilon_0)e^x, \\ Q &= \mu + \gamma\cos y, \quad \epsilon_0 = \{1,0,-1\}, \end{aligned} \tag{4.7}$$

establishing in this manner that the studied solution is certainly endowed with three arbitrary continuous parameters.

The metric (4.7) for vanishing parameter μ reduces, modulo minor redefinitions, to the Kramer fluid solution, see (16) in Ref. 2.

The obtained twisting and nondiverging metric can also be written as

$$\begin{aligned} g = {}& (e^{2\nu x}/P)dx^2 + (e^{2\nu x}/Q)dy^2 \\ & + \kappa^2 e^{-2\nu x}[P - Q(1 - e^{2\nu(x-x_0)})^2]d\sigma^2 \\ & + 2\kappa e^{-2\nu x}[P - Q(1 - e^{2\nu(x-x_0)})]d\tau\,d\sigma \\ & - e^{-2\nu x}(Q-P)d\tau^2, \\ P = {}& \mu + \delta(x+\epsilon_0)e^{2\nu x}, \quad Q = \mu + \gamma\cos 2\nu y, \\ \epsilon_0 = {}& \{1,0,-1\}, \\ \kappa\delta = {}& \pm 2[1 + 2\nu(\epsilon_0 + x_0)]^{-1}\exp(-\nu x_0). \end{aligned} \tag{4.8}$$

Thus the coordinate τ has to be interpreted as the time coordinate, while σ is interpretable as the azimuthal coordinate. The x_0 value of x, which is defined as a solution of the equation $P(x_0) = 0$ determines the axis of symmetry and rotation. One can easily establish that (4.8) satisfy at the rotation axis x_0 the regularity condition (2.10).

Defining the fluid one-form $u = u_a e^a$, where $u_a = u_\mu e^\mu_a$ are the tetrad components of the fluid four-velocity (2.2), one obtains

$$u = \frac{i}{\sqrt{2}}\left(\frac{P}{Q-P}\right)^{1/2}(e^2 - e^1) + \frac{1}{\sqrt{2}}\left(\frac{Q}{Q-P}\right)^{1/2}(e^4 - e^3), \tag{4.9}$$

therefore, the three-form

$$u \wedge e^3 \wedge e^4 = (i/\sqrt{2})\left(\frac{P}{(Q-P)}\right)^{1/2}(e^2 - e^1)\wedge e^3 \wedge e^4 \neq 0. \tag{4.10}$$

Thus u does not lie in the two-space spanned by the null principal directions e^3 and e^4. Consequently, the obtained solution belongs to class II of the Wainwright classification.[8]

V. FORMAL TWIST-FREE AND DIVERGING SOLUTION

The third case of our schema, for which $M_x = 0 \neq N_y$ yields a three-parameter solution to the Einstein perfect fluid equations possessing negative energy density, and therefore of little physical interest. Nevertheless, from the completeness viewpoint we consider it pertinent to present it here. This twist-free formal solution is given by the metric (2.1) with structural functions,

$$\begin{aligned} & P = \mu + \gamma\cos 2\nu x, \quad Q = \mu + (ay + b)e^{2\nu y}, \\ & M = l, \quad N = l - \kappa e^{2\nu y}, \quad \Delta := \kappa e^{2\nu y}, \end{aligned} \tag{5.1}$$

where μ, ν, γ, κ, a, b, and l are integration constants.

The energy density and the pressure of the fluid for this solution are given by

$$\begin{aligned} \epsilon &= -3(\nu^2/\Delta)(Q-P) - \nu a/\kappa\,, \\ p &= (\nu^2/\Delta)(Q-P) + \nu a/\kappa\,, \\ \epsilon + 3p &= \nu a/\kappa\,. \end{aligned} \tag{5.2}$$

Since in the zero pressure surface $p = 0$, the energy reduces to $\epsilon_s = \nu a/\kappa$, which has to be positive ($\epsilon_s > 0$), then energy density ϵ from (5.2) becomes a negative quantity.

VI. LIMITING CONTRACTION

In this section we shall derive twisting and divergenceless solution as limiting contractions of the twisting and diverging (Wahlquist) metric. In order to accomplish the transition, we introduce a contraction parameter ϵ and subject the metric (2.1) to the coordinate transformations

$$x \to \epsilon^{-1}(x + (1/2\nu)\ln 2\epsilon), \quad y \to y\epsilon^{-1}, \quad \tau \to \epsilon\tau', \quad \sigma \to \sigma, \tag{6.1}$$

accompanied by a redefinition of the parameters appearing in the structural functions M and N from (3.2) and P and Q from (3.8) according to

$$\begin{aligned} & \nu \to \epsilon\nu, \quad \kappa \to \kappa, \quad \beta \to \nu a, \quad a \to \mu\epsilon^{-1}, \quad \mu \to \epsilon l, \quad b \to \gamma\epsilon^{-1}, \\ & m \to m\epsilon^{-2}, \quad n \to \gamma\epsilon^{-1} + (a/2\nu)\epsilon^{-2}\ln 2\epsilon - b\epsilon^{-2}. \end{aligned} \tag{6.2}$$

In the limit $\epsilon \to \infty$, one establishes that

$$\lim_{\epsilon\to\infty}\{g, M\epsilon^{-1}, N\epsilon^{-1}, P\epsilon, Q\epsilon\}_{\mathrm{WS}} \to \{g, M, N, P, Q\}_{\mathrm{DS}}, \tag{6.3}$$

where the subscripts WS and $\not{D}$S denote Wahlquist and divergenceless solutions, respectively.

ACKNOWLEDGMENTS

We are grateful to Consejo Nacional de Ciencia y Tecnológia (CONACyT) and Departamento de Física del CINVESTAV for the financial assistence and hospitality. This paper was also partially supported by COSNET-SEP.

[1] H. D. Wahlquist, Phys. Rev. **172**, 1291 (1968).
[2] D. Kramer, Class. Quantum Gravit. **1**, L3 (1984).
[3] B. Carter, Commun. Math. Phys. **10**, 280 (1968).
[4] S. Bonanos, Commun. Math. Phys. **19**, 53 (1976).
[5] We owe the referee for the quoted information.
[6] D. Kramer, Astron. Nachr. **307**, 309 (1986).
[7] P. C. Vaidya, Pramana **8**, 512 (1977).
[8] J. Wainwright, Gen. Relativ. Gravit. **8**, 797 (1977).

Exact solutions of the Einstein field equations for a topologically nontrivial metric

Tina A. Harriott
Department of Mathematics, Mount Saint Vincent University, 166 Bedford Highway, Halifax, Nova Scotia B3M 2J6, Canada

J. G. Williams
Department of Mathematics and Computer Science, Brandon University, Brandon, Manitoba R7A 6A9, Canada

(Received 12 August 1987; accepted for publication 2 September 1987)

Two exact solutions of the Einstein field equations are presented, each having a Finkelstein–Misner kink. The first of these has a perfect fluid interior and a vacuum exterior. The equation of state is $\rho = -p = C$ (where C is a constant), which is the same as that found in inflationary models.

I. INTRODUCTION

In a previous paper[1] the present authors discussed the kink metric[2]

$$g_{\mu\nu} = \delta_{\mu\nu} - 2\phi_\mu \phi_\nu$$

and noted the following relationship:

$$\phi^\mu \phi_\mu = g^{\mu\nu}\phi_\nu \phi_\mu = g_{\mu\nu}\phi^\nu \phi^\mu = -1 .$$

Because of the timelike behavior of ϕ^μ, it was proposed that the ϕ^μ be interpreted as the components of the four-velocity of a fluid. The functions ϕ_μ (and also the functions $\phi^\mu = g^{\mu\nu}\phi_\nu = -\phi_\mu$) take values on a three-sphere:

$$\sum \phi_\mu \phi_\mu = \sum \phi^\mu \phi^\mu = 1 .$$

Thus at any instant of time, the $\{\phi_\mu\}$ represent a mapping $\varphi:\mathbb{R}^3 \to S^3$. As $|x| \to \infty$, one obtains asymptotic flatness $(g_{\mu\nu} \to \eta_{\mu\nu})$ by imposing the condition $(\phi^0,\phi^1,\phi^2,\phi^3) \to (1,0,0,0)$. The degree of the mapping φ then equals the number of Finkelstein–Misner kinks[3] present in the metric $g_{\mu\nu}$.

The purpose of the present paper is to seek one-kink solutions of the Einstein equations using the above metric, but with the $\{\phi^\mu\}$ given in terms of the Skyrme hedgehog[4] according to

$$\phi^0 = \cos\alpha, \quad \phi^i = (x^i/r)\sin\alpha, \quad i = 1,2,3 .$$

The angle α is a function only of $r = (x^2+y^2+z^2)^{1/2}$ and for a one-kink metric (i.e., a degree one map onto S^3) we require[4] α to be continuous and to satisfy

$$\alpha(0) = \pi, \quad \alpha(\infty) = 0.$$

Since α is known[5] to equal the angle of tilt of the light cone, one can see from the form of the hedgehog that, as the cones tip over on the way from infinity to the origin, the fluid velocity is confined to directions that lie within each of the cones, as is required for a vector describing movement of physical material.

The hedgehog assumption leads to the following (spherically symmetric) form for the metric:

$$g_{00} = -\cos 2\alpha, \quad g_{0i} = -(x^i/r)\sin 2\alpha,$$
$$g_{ij} = (x^i x^j/r^2)\cos 2\alpha + \tau_{ij} ,$$

where τ_{ij} is defined by

$$\tau_{ij} = \delta_{ij} - x^i x^j/r^2$$

and satisfies

$$\tau_{ij}x^j = 0, \quad \sum \tau_{ii} = 2, \quad \sum \tau_{ij}\tau_{jk} = \tau_{ik} .$$

Even though the metric is spherically symmetric, the time-space term cannot be (globally) removed by a coordinate transformation. The usual procedure[6] for doing this involves defining a new time coordinate $\bar{t} = t + f(r)$. Working in spherical polar coordinates, the components g_{0r} transform according to

$$\bar{g}_{0r} = g_{0r} + g_{00}\frac{\partial t}{\partial \bar{r}} = g_{0r} - g_{00}\frac{df}{dr} .$$

Hence choosing the function f so that $df/dr = g_{0r}/g_{00}$ should remove the time-space term (in the new coordinates). However, with a kink present, such a choice of f is not globally possible since g_{00} will equal zero at least once somewhere.

Such singular transformations have recently been discussed by Rosen,[7] who presents a number of inequivalent spherically symmetric vacuum solutions of the Einstein equations. Rosen[7] points out that two metrics that can be transformed into each other *only by transformations that have singularities* should not be regarded as the same metric and should not be thought of as describing the same physical situation. Any transformation that can remove the kink from the above metric $g_{\mu\nu}$ will be singular at least at one point and will be regarded as inadmissible for the purposes of the present paper.

Because of the nonzero g_{0i} term, the metric $g_{\mu\nu}$ is not static. In fact, $g_{\mu\nu}$ is not even stationary in the sense that it does not have a *globally* timelike Killing vector. The computation of the Killing vectors[8] yields the usual three Killing vectors appropriate for spherical symmetry and a Killing vector ξ of magnitude $(-\cos 2\alpha)$ that changes from timelike to spacelike as α varies.

In what follows, we shall determine the unknown function $\alpha(r)$ so that the Einstein equations $G_{\mu\nu} = 8\pi T_{\mu\nu}$ are satisfied. The $(-+++)$ convention of Misner, Thorne,

0022-2488/88/010179-03$02.50

and Wheeler[9] will be followed throughout. The symbol $\delta_{\mu\nu}$ is used to denote the components of the Kronecker delta, whereas $\eta_{\mu\nu}$ denotes the components of the Minkowski metric. The topology of the space-time manifold $\mathscr{M}$ is assumed to be trivial, $\mathscr{M} = \mathbb{R}^4$.

II. CURVATURE AND KINEMATICS

The Christoffel symbols and Ricci and Einstein tensors for the above metric are readily obtained from the formulas given by Harriott and Williams[1] and are also listed in Williams and Zia[2] and Finkelstein and McCollum.[5] (For an alternative approach related to the present work, see Clément.[10]) Since $\det g = -1$, it follows that $g^{\mu\nu} = g_{\mu\nu}$ and $\Gamma^{\mu}_{\mu\nu} = 0$, for all ν. It is straightforward to show that

$$G^0_0 = (-2/r^2)\partial_r(r\sin^2\alpha), \quad G^i_0 = G^0_i = 0,$$

$$G^j_i = G^0_0 x^i x^j/r^2 + (\tfrac{1}{2}r)\partial_r(r^2 G^0_0)\tau_{ij},$$

$$R = (2/r^2)\partial^2_r(r^2\sin^2\alpha).$$

Following Ellis,[11] we note that the stress-energy tensor $T^{\ \nu}_{\mu}$ can be written as

$$T^{\ \nu}_{\mu} = \rho u_\mu u^\nu + q_\mu u^\nu + u_\mu q^\nu + p h^{\ \nu}_{\mu} + \pi^\nu_\mu,$$

where q_μ is the energy flux (due to diffusion/heat conduction), $h^{\ \nu}_{\mu}$ is the projection tensor (equal to $\delta^\nu_\mu + u_\mu u^\nu$), π^ν_μ is the anisotropic pressure (viscosity) term, and $q_\mu u^\mu = \pi^\mu_\mu = \pi^\nu_\mu u_\nu = 0$. It is usual[11] to assume that π^ν_μ is linearly related to the shear tensor σ^ν_μ:

$$\pi^\nu_\mu = -\lambda\sigma^\nu_\mu,$$

where the constant λ is positive and is proportional to the coefficient of (dynamic) viscosity.

The metric, together with the above choice of fluid velocity, namely, $u^0 = \cos\alpha$, $u^i = (x^i/r)\sin\alpha$, can then be used to evaluate the various quantities occurring in the expression for $T^{\ \nu}_{\mu}$. The components of the projection tensor $h^{\ \nu}_{\mu}$ are

$$h^{\ 0}_{0} = \sin^2\alpha, \quad h^{\ i}_{0} = h^{\ 0}_{i} = -(x^i/r)\sin\alpha\cos\alpha,$$

$$h^{\ j}_{i} = \delta^j_i - (x^i x^j/r^2)\sin^2\alpha.$$

The shear tensor is defined in terms of the covariant derivatives $u_{\mu;\nu}$ and the isotropic (volume) expansion $\theta \equiv u^\mu_{;\mu}$:

$$\sigma_{\mu\nu} \equiv (u_{\mu;\eta}h^{\ \eta}_{\nu} + u_{\nu;\eta}h^{\ \eta}_{\mu})/2 - \theta h_{\mu\nu}/3.$$

The $u_{\mu;\nu}$ and θ are given by

$$u_{0;0} = \sin 2\alpha \sin\alpha\,\partial_r\alpha,$$

$$u_{i;0} = -(x^i/r)\sin 2\alpha\cos\alpha\,\partial_r\alpha,$$

$$u_{0;j} = -(x^j/r)\cos 2\alpha\sin\alpha\,\partial_r\alpha,$$

$$u_{i;j} = (x^i x^j/r^2)\cos 2\alpha\cos\alpha\,\partial_r\alpha + (\tau_{ij}/r)\sin\alpha,$$

$$\theta = r^{-2}\partial_r(r^2\sin\alpha).$$

The acceleration vector $\dot{u}^\mu = u^\mu_{;\nu}u^\nu$ is

$$(\dot{u}^0, \dot{u}^i) = (\dot{u}_0, \dot{u}_i) = (\sin\alpha, -(x^i/r)\cos\alpha)\sin\alpha\,\partial_r\alpha.$$

The shear tensor σ^ν_μ and the scalar shear $\sigma = (\sigma^{\mu\nu}\sigma_{\mu\nu})^{1/2}$ are as follows:

$$\sigma = 3^{-1/2} r\,\partial_r(r^{-1}\sin\alpha), \quad \sigma^0_0 = \sigma 3^{-1/2}(1 - \cos 2\alpha),$$

$$\sigma^i_0 = \sigma^0_i = -\sigma 3^{-1/2}(x^i/r)\sin 2\alpha,$$

$$\sigma^j_i = \sigma 3^{-1/2}[(x^i x^j/r^2)(1 + \cos 2\alpha) - \tau_{ij}].$$

Since the chosen metric is spherically symmetric, it is clear that the vorticity tensor $\omega^\nu_\mu \equiv (u_{u;\eta}h^{\ \eta}_{\mu} - u_{\nu;\eta}h^{\ \eta}_{\mu})/2$ will be zero everywhere. We also remark that $q_\mu u^\mu = 0$ leads to the following relationship between the components of the heat flux vector q_μ:

$$q_i(x^i/r)\sin\alpha = -q_0\cos\alpha.$$

III. EINSTEIN EQUATIONS

Henceforth, π^ν_μ will be equated to $-\lambda\sigma^\nu_\mu$, where λ is a positive constant. Then $G^i_0 = 8\pi T^i_0$ leads to

$$(\rho + p - 2\lambda\sigma/3^{1/2})u_0 u^i + q_0 u^i + u_0 q^i = 0.$$

If functions f_1 and f_2 are defined by $q_0 = f_1 u_0$, $q^i = f_2 u^i$, then the term $q_0 u^i + u_0 q^i$ can be written as $(f_1 + f_2)u_0 u^i$. The condition $q_\mu u^\mu = 0$ now implies $f_1 u_0 u^0 + f_2 u_i u^i = 0$, so that $f_1\cos^2\alpha + f_2\sin^2\alpha = 0$. Hence $f_1 = -f\sin^2\alpha$ and $f_2 = f\cos^2\alpha$, where f is an as yet unspecified (and possibly zero) function. Thus

$$q^0 = -f\sin^2\alpha\cos\alpha, \quad q^i = (x^i/r)f\sin\alpha\cos^2\alpha.$$

If any one of f_1, f_2, or f are zero, then all three will be zero. We now have the following relation:

$$\rho + p - 2\lambda\sigma/3^{1/2} + f_1 + f_2 = 0.$$

Then $G^j_i = 8\pi T^j_i$ gives

$$G^0_0\, x^i x^j/r^2 + (1/2r)\partial_r(r^2 G^0_0)\tau_{ij}$$

$$= 8\pi(\rho u_i u^j + p u_i u^j + p\delta^j_i - \lambda\sigma^j_i + q_i u^j + u_i q^j)$$

$$= 8\pi[(\rho + p - 2\lambda\sigma/3^{1/2} + 2f_2)u_i u^j$$

$$- (\lambda\sigma/3^{1/2})\{(2x^i x^j/r^2) - \tau_{ij}\} + p\delta^j_i].$$

Comparison of the above equations suggests that we choose f_1 and f_2 to be equal everywhere; it then follows from their definitions that they must both equal zero. From now on, we assume $f_1 = f_2 = 0$, which leads to the following equation of state relating energy density, pressure, and shear:

$$\rho + p - 2\lambda\sigma/3^{1/2} = 0.$$

Then $G^j_i = 8\pi T^j_i$ becomes

$$G^0_0 x^i x^j/r^2 + (1/2r)\partial_r(r^2 G^0_0)\tau_{ij}$$

$$= 8\pi[p\delta^j_i - (\lambda\sigma/3^{1/2})\{(2x^i x^j/r^2) - \tau_{ij}\}]$$

and $G^0_0 = 8\pi T^0_0$ becomes $G^0_0 = 8\pi(p - 2\lambda\sigma/3^{1/2})$. (One may check from the equation of state that $T^0_0 = -\rho$.)

IV. PERFECT FLUID SOLUTION

For a perfect fluid, the stress-energy tensor is $T^{\ \nu}_{\mu} = (\rho + p)u_\mu u^\nu + p\delta^\nu_\mu$, with π^ν_μ absent and $\lambda = 0$. The equation of state given previously now becomes $\rho = -p$. The field equations of interest are

$$G^0_0 x^i x^j/r^2 + (1/2r)\partial_r(r^2 G^0_0)(\delta_{ij} - x^i x^j/r^2) = 8\pi p\delta^j_i,$$

$$G^0_0 = (-2/r^2)\partial_r(r\sin^2\alpha) = 8\pi p;$$

the first of these can be solved by choosing G^0_0 (and hence the pressure p) to be a constant. If the constant is assumed nonzero, $p = -C \neq 0$, the second equation leads to an interior solution

$$\sin^2\alpha = (4\pi/3)Cr^2,$$

which requires C to be positive and hence the energy density $\rho = -p = C$ to be a positive constant. Fluids satisfying this relation have arisen in inflationary cosmological models[12] and in certain particle models.[13] Alternatively, choosing $\rho = -p = 0$ leads to an exterior solution

$$\sin^2\alpha = M/r,$$

when M is a positive constant. For a kink to be present, these two solutions must be joined so that $\alpha(0) = \pi$, $\alpha(\infty) = 0$, with $\sin\alpha$ rising from 0 to 1 and failing back to 0 again as r increases from the origin to infinity. This is achieved by the following overall solution:

$$\sin\alpha = \begin{cases} r/M, & 0 \leqslant r \leqslant M, \\ (M/r)^{1/2}, & M \leqslant r < \infty, \end{cases}$$

with $M = (3/4\pi C)^{1/2}$. The solution describes an object of radius M surrounded by empty space. It is easy to check that M is the total mass. The exterior solution is not Schwarzschild since it is not possible to transform away the g_{0i} term. To transform the external solution into the Schwarzschild form would require "unfastening" the solution from the boundary at $r = M$. These, of course, are *global* comments. The exterior solution is *locally* transformable into the Schwarzschild solution, in accordance with Birkhoff's theorem. It should be remarked that Rosen[7] has given a number of spherically symmetric exterior vacuum solutions that are not globally transformable in the Schwarzschild solution (in a *nonsingular* way). Unlike the solutions given in the present paper, Rosen's[7] solutions are not kinked.

For the interior solution, the scalar curvature is $R = 32\pi C$ and there is a positive expansion $\theta = 3/M$. The scalar shear is found to be zero, $\sigma = 0$, so that $\sigma_{\mu\nu} = 0$ for all μ,ν.

In the exterior region, the scalar curvature is zero, $R = 0$, but θ and σ are found to be nonzero. (Since $\rho = 0$, the interpretation of nonzero θ and σ is unclear, but models with such behavior have arisen in other situations.[8])

V. IMPERFECT FLUID SOLUTION

The field equations show that the only interior solution is the previously chosen ($\lambda = 0$) perfect fluid solution. With $\lambda > 0$, the field equations lead to

$$G^0_0 - (1/2r)\partial_r(r^2 G^0_0) = -24\pi\lambda\sigma/3^{1/2},$$

$$(1/2r)\partial_r(r^2 G^0_0) = 8\pi(p + \lambda\sigma/3^{1/2}),$$

$$G^0_0 = (-2/r^2)\partial_r(r\sin^2\alpha)$$

$$= 8\pi(p - 2\lambda\sigma/3^{1/2}) = -8\pi\rho.$$

The following may be shown to be an exterior solution:

$$\sin\alpha = -(\lambda/3)r + M^{1/2}(1 + \lambda M/2)r^{-1/2}.$$

The outer edge of this solution must be joined to the trivial solution $\sin\alpha \equiv 0$ to preserve the boundary condition at infinity. This solution will be pursued no further, except to note that $\rho = 2\lambda^2/3 + 4\lambda\sigma/3^{1/2}$ is not positive everywhere.

VI. CONCLUSIONS

We have demonstrated the existence of exact solutions for the metric $g_{\mu\nu} = \delta_{\mu\nu} - 2\phi_\mu\phi_\nu$. The most interesting of these was a perfect fluid solution with an exterior vacuum. The interior was similar to an inflation metric, having constant energy density $\rho = -p$. The expansion factor is exponential, $L \propto e^{t/M}$, since $\dot{L}/L = \theta/3 = M^{-1}$. Starting from[11] $q_\mu = -\kappa h^{\nu}_{\mu}(T_{;\nu} + T\dot{u}_\nu)$ (with $q_\mu = 0$), it can be shown[8] that the temperature falls off exponentially with increasing r. The manifold is well-behaved everywhere,[8] including $r = 0$, so that there are no curvature singularities.

ACKNOWLEDGMENTS

We would like to thank Dr. K. A. Dunn for his interest in this work and the referee for a number of helpful comments. Useful discussions with Dr. David Finkelstein and other participants of the Quantum Topology Workshop at Georgia Institute of Technology are gratefully acknowledged.

One of us (JGW) would like to thank the Natural Sciences and Engineering Research Council of Canada for financial support through operating Grant No. A9151.

[1] T. A. Harriott and J. G. Williams, J. Math. Phys. **27**, 2706 (1986).

[2] J. G. Williams and R. K. P. Zia, J. Phys. A **6**, 1 (1973).

[3] D. Finkelstein and C. W. Misner, Ann. Phys. (NY) **6**, 230 (1959).

[4] T. H. R. Skyrme, Proc. R. Soc. London Ser. A **260**, 127 (1961).

[5] D. Finkelstein and G. McCollum, J. Math. Phys. **16**, 2250 (1975).

[6] P. G. Bergmann, *Introduction to the Theory of Relativity* (Prentice–Hall, Englewood Cliffs, NJ, 1959), Chap. 13.

[7] N. Rosen, Found. Phys. **15**, 517 (1985).

[8] T. A. Harriott, Ph.D. thesis, Dalhousie University, 1987 (unpublished).

[9] C. W. Misner, K. S. Thorne, and J. A. Wheeler, *Gravitation* (Freeman, San Francisco, 1973).

[10] G. Clément, Gen. Relativ. Gravit. **16**, 131, 477, 491 (1984); **18**, 137 (1986).

[11] G. F. R. Ellis, "Relativistic cosmology," in *Proceedings of the International School of Physics "Enrico Fermi," Course XLVII–General Relativity and Cosmology*, Varenna, Lake Como, 30 June–12 July 1969, edited by B. K. Sachs (Academic, New York, 1971).

[12] A. H. Guth, Phys. Rev. D **23**, 347 (1981); "The new inflationary universe 1984," in *Innerspace and Outerspace*, edited by E. W. Kolb, M. S. Turner, D. Lindley, K. Olive, and D. Seckel (Univ. Chicago, Chicago, 1986).

[13] N. Rosen, "The field of a particle in general relativity," in *Unified Field Theories of More Than 4 Dimensions–Eighth Course of the International School of Cosmology and Gravitation of the Ettore Majorana International Centre for Scientific Culture*, Erice, 1982, edited by V. De Sabbata and E. Schmutzer (World Scientific, Singapore, 1983).

On the Kerr–Newman metric in cosmological background

Sharda S. Koppar and L. K. Patel
Department of Mathematics, Gujarat University, Ahmedabad 380 009, India

(Received 15 April 1986; accepted for publication 5 August 1987)

Using the method of null tetrad, the Kerr–Newman metric in the background of the Robertson–Walker universe is derived in terms of a metric that is conformal to a generalized Kerr–Schild metric. A new two–fluid interpretation of this metric is presented. The Kerr–Newman metric in the background of the Einstein–de Sitter universe is discussed in detail.

I. INTRODUCTION

The problem of finding exterior gravitational fields of black holes embedded in some world models has attracted wide attention. Vaidya[1] has discussed the exterior gravitational field of a Kerr[2] black hole embedded in the Robertson–Walker universe with positive curvature of the spacelike surfaces $t = \text{const}$. The source for the above solution is an anisotropic fluid (i.e., pressures in all three spatial directions are not equal). Taub[3] has shown that the source for Vaidya's solution can be taken as a mixture of perfect fluid and a null fluid. Patel and Trivedi[4] have generalized Vaidya's solution to include source-free electromagnetic fields. Their solution describes the Kerr–Newman[5] metric in the background of the closed Robertson–Walker universe. The source of their solution is also an anisotropic fluid.

We know that the method of null tetrads is widely used in solving the problems of relativity theory. Therefore it would be interesting to rederive the Kerr–Newman metric in the cosmological background by this method.

The object of the present investigation is to give this rederivation and to give a new two-fluid interpretation of the Kerr–Newman metric in the cosmological background of the closed Robertson–Walker universe. We also intend to discuss the Kerr–Newman solution in the background of the Einstein–de Sitter universe.

II. THE METRIC AND THE EINSTEIN TENSOR

A space-time with metric tensor $\hat{g}_{ik}$ will be said to be a generalized Kerr–Schild space-time $\hat{V}$ when

$$\hat{g}_{\mu\gamma} = g_{\mu\gamma} + 2Hl_\mu l_\gamma, \tag{2.1}$$

where $g_{\mu\gamma}$ is the metric of an arbitrary space-time V, H is a scalar field over V, and l_μ is a null, geodesic, and shear-free vector field in V. Let us consider a space-time V^* that is conformal to V. Taub[3] has verified that l_μ remains null, geodesic, and shear-free in $\hat{V}$ and V^*.

In the present paper we shall take V to be the Einstein universe. Vaidya[1] has shown that the metric of the Einstein universe can be expressed in the form

$$ds^2 = -dt^2 + dr^2 + (|\rho|^2/N^2)d\alpha^2 + (|\rho|^2 + k^2 \sin^2\alpha) \times \sin^2\alpha\, d\beta^2 - 2k \sin^2\alpha\, d\beta\, dr, \tag{2.2}$$

where

$$\rho = (R_0^2 - k^2)^{1/2} \sin(r/R_0) + ik \cos\alpha,$$
$$N^2 = 1 - (k^2/R_0^2)\sin^2\alpha.$$

Here R_0 and k are constants. Note that the cosmological constant Λ is nonzero for the Einstein universe.

It may be verified that the four null vectors l_μ, n_μ, m_μ, and $\overline{m}_\mu$ given by the equations

$$\begin{aligned} \sqrt{2}l_\mu &= (-1,0,k\sin^2\alpha,-1),\\ \sqrt{2}n_\mu &= (1,0-k\sin^2\alpha,-1),\\ \sqrt{2}m_\mu &= (0,\rho/N,i\rho\sin\alpha,0),\\ \sqrt{2}\overline{m}_\mu &= (0,\bar{\rho}/N,-i\bar{\rho}\sin\alpha,0), \end{aligned} \tag{2.3}$$

are such that the metric tensor of (2.2) becomes

$$g_{\mu\gamma} = -(l_\mu n_\gamma + l_\gamma n_\mu) + m_\mu \overline{m}_\gamma + m_\gamma \bar{m}_\mu. \tag{2.4}$$

Here and in what follows an overbar indicates a complex conjugate. Also these vectors satisfy

$$l^\mu n_\mu = -1, \quad m^\mu \overline{m}_\mu = 1. \tag{2.5}$$

All other inner products are 0. From (2.2) and (2.3) we can find the contravariant components of these vectors. They are given by

$$\begin{aligned} \sqrt{2}l^\mu &= (-1,0,0,1),\\ \sqrt{2}n^\mu &= (1,0,0,1),\\ \sqrt{2}m^\mu &= (i\rho/|\rho|^2)(k\sin\alpha,-iN,\csc\alpha,0),\\ \sqrt{2}\overline{m}^\mu &= -(i\bar{\rho}/|\rho|^2)(k\sin\alpha,iN,\csc\alpha,0). \end{aligned} \tag{2.6}$$

If we define a vector field U_μ by $\sqrt{2}U_\mu = l_\mu + n_\mu$ then the following results about U_μ can be easily established:

$$U_\mu l^\mu = U_\mu n^\mu = -1/\sqrt{2}, \quad U^\mu U_\mu = -1, \; U_{\mu|\gamma} = 0. \tag{2.7}$$

The stroke denotes the covariant derivative with respect to $g_{\mu\gamma}$. It is a straightforward matter to verify that

$$l_{\mu|\gamma} m^\mu n^\gamma = \theta. \tag{2.8}$$

If θ and Ω are the expansion and the rotation of the null vector l_μ, then $Z = \theta - i\Omega = l_{\mu|\gamma} m^\mu \overline{m}^\gamma$. Using (2.3) and (2.6) it can be seen that

$$|\rho|^2 Z = -\sqrt{2}\left[\frac{(R_0^2 - k^2)}{R_0^2} \sin\left(\frac{r}{R_0}\right)\cos\left(\frac{r}{R_0}\right) - ikN\cos\alpha\right]. \tag{2.9}$$

The Ricci tensor and the scalar curvature for the Einstein universe are given by (see Taub[3])

0022-2488/88/010182-04$02.50

$$R_{\mu\gamma} = -(2/R_0^2)(g_{\mu\gamma} + U_\mu U_\gamma), \quad R = -(6/R_0^2). \tag{2.10}$$

The following relations can be established by a straightforward computation:

$$\begin{aligned} &Z_{,\mu} l^\mu = -Z_{,\mu} n^\mu = -(1/R_0^2 + Z^2/2), \\ &Z_{,\mu} m^\mu = 0, \\ &Z_{,\mu} \overline{m}^\mu = \frac{ik \sin \alpha \overline{\rho}}{|\rho|^2} \bigg[-Z^2 + \frac{Z\overline{Z}}{2} - \frac{1}{R_0^2} \\ &\qquad + \frac{NN_\alpha \cot \alpha - N^2}{|\rho|^2} \bigg], \end{aligned} \tag{2.11}$$

where the comma indicates partial derivative and $N_\alpha = \partial N / \partial \alpha$.

We now consider a space-time V^* with the metric tensor

$$g^*_{\mu\gamma} = S^2(t)[g_{\mu\gamma} + 2Hl_\mu l_\gamma], \tag{2.12}$$

where $g_{\mu\gamma}$ is the metric tensor of the Einstein universe described by (2.2), S is an arbitrary function of t, and H is a scalar function of coordinates.

The expression for Einstein tensor for the space-time V^* has been worked out by Taub.[3] We state this expression for ready reference:

$$\begin{aligned} G^{*J}_\gamma = {} & X(g^{\mu J} - 2Hl^\mu l^J) U_\mu U_\gamma - N^* l^J l_\gamma \\ & + [X/2(1+H) - P^* - (\Gamma_0 + E)/S^4] \delta^J \gamma \\ & + \pi^* (\overline{m}^J m_\gamma + m^J \overline{m}_\gamma) - \Delta(\overline{m}^J l_\gamma + \overline{m}_\gamma l^J) \\ & - \overline{\Delta}(m^J l_\gamma + m_\gamma l^J), \end{aligned} \tag{2.13}$$

where various quantities are given by

$$\begin{aligned} & S^2 X = 2(SS^{11} - S^{12} - 1/R_0^2), \\ & S^2 P^* = -(SS^{11} + 2S^{12} + 2/R_0^2), \\ & -S^4 N^* = -N_0 + 2\sqrt{2}(H_0 n^\rho)_{|\rho} S^1 + 4H_0/R_0^2, \\ & S^4 \pi^* = \Phi_{0\rho} l^\rho - \Gamma_0 - E, \\ & S^4 \Delta = \Delta_0, \\ & \Gamma_0 = \tfrac{1}{2}(H_0 l^\rho)_{|\rho}(Z + \overline{Z}) - \tfrac{1}{2} H_0 Z\overline{Z} + H_0 R_{\mu\gamma} l^\mu l^\gamma, \\ & E = H_0 SS^{11} + \sqrt{2} S^1 (H_0 l^\rho)_{|\rho}, \\ & N_0 = -H_{0|\mu\gamma} g^{\mu\gamma} - 2\Phi_{0\rho} n^\rho, \ \Delta_0 = \Phi_{0\gamma} m^\gamma, \\ & \Phi_{0\mu} = (H_0 l^\rho)_{|\rho\mu} - (H_0 l_{\mu|\rho\sigma} + 2H_{0,\sigma} l_{\mu|\rho}) g^{\rho\sigma} \\ & \qquad - H_0 l^\sigma R_{\sigma\mu}. \end{aligned} \tag{2.14}$$

Here

$$H_0 = HS^2. \tag{2.15}$$

It should be noted that we have used the results (2.1)–(2.12) in arriving at the expression (2.13) of the Einstein tensor. An overbar indicates differentiaton with respect to t.

III. THE ELECTROMAGNETIC FIELD

We choose the electromagnetic four-potential A^*_μ as

$$A^*_\mu = \phi l^*_\mu, \tag{3.1}$$

where $l^*_\mu = l_\mu$ and ϕ is a function of coordinates.

Therefore

$$l^{*\mu} = l^\mu / S^2. \tag{3.2}$$

The electromagnetic field tensor corresponding to the choice (3.1) is given by

$$F^*_{\mu\gamma} = \phi(l_{\mu|\gamma} - l_{\gamma|\mu}) + \phi_{,\gamma} l_\mu - \phi_{,\mu} l_\gamma, \tag{3.3}$$

$$F^{*\mu\gamma} = S^{-4}[\phi(l^{\mu|\gamma} - l^{\gamma|\mu}) + \phi^{,\gamma} l^\mu - \phi^{,\mu} l^\gamma]. \tag{3.4}$$

Let us now try to solve the source-free Maxwell equations $F^{*\mu\gamma}_{;\gamma} = 0$, where the semicolon indicates the covariant derivative with respect to the metric tensor $g^*_{\mu\gamma}$. The relations $F^{*\mu\gamma}_{;\gamma} l_\mu = 0$, $F^{*\mu\gamma}_{;\gamma} m_\mu = 0$, and $F^{*\mu\gamma}_{;\gamma} \overline{m}_\mu = 0$ imply the following equations:

$$\phi_{,\mu\gamma} l^\mu l^\gamma + \phi_{,\mu} l^\mu \left(\frac{Z + \overline{Z}}{2}\right) + \phi \left(\frac{Z - \overline{Z}}{2}\right)^2 = 0, \tag{3.5}$$

$$\begin{aligned} & Z\phi_{,\mu} m^\mu - \tfrac{1}{2}\phi \overline{Z}_{,\mu} m^\mu - (\phi_{,\gamma} l^\gamma)_{|\mu} m^\mu \\ & \qquad - \phi_{,\mu} m^\mu [(Z + \overline{Z})/2] = 0, \end{aligned} \tag{3.6}$$

$$\begin{aligned} & \overline{Z}\phi_{,\mu} \overline{m}^\mu - \tfrac{1}{2}\phi Z_{,\mu} \overline{m}^\mu - (\phi_{,\gamma} l^\gamma)_{|\mu} \overline{m}^\mu \\ & \qquad - \phi_{,\mu} \overline{m}^\mu [(\overline{Z} + Z)/2] = 0. \end{aligned} \tag{3.7}$$

Equation (3.5) will determine ϕ as

$$\phi = \tfrac{1}{2} Q(Z + \overline{Z}), \quad Q_{,\mu} l^\mu = 0. \tag{3.8}$$

Substituting ϕ from (3.8) in (3.6) and (3.7) we get

$$Q_{,\mu} m^\mu = 0, \quad Q_{,\mu} \overline{m}^\mu = 0. \tag{3.9}$$

Since Q is required to be real and it satisfies $Q_{,\mu} l^\mu = 0$, $Q_{,\mu} m^\mu = 0$, $Q_{,\mu} \overline{m}^\mu = 0$, the integrability conditions of these equations imply that we must have

$$\begin{aligned} & (l^\gamma m^\mu_{|\gamma} - m^\gamma l^\mu_{|\gamma}) Q_{,\mu} = 0, \\ & (l^\gamma \overline{m}^\mu_{|\gamma} - \overline{m}^\gamma l^\mu_{|\gamma}) Q_{,\mu} = 0, \\ & (\overline{m}^\gamma m^\mu_{|\gamma} - m^\gamma \overline{m}^\mu_{|\gamma}) Q_{,\mu} = 0. \end{aligned}$$

These conditions can be expressed as $Q_{,\mu} n^\mu (Z - \overline{Z}) = 0$. As $Z \neq \overline{Z}$ (i.e., $\Omega \neq 0$), we have $Q_{,\mu} n^\mu = 0$. Thus Q is a constant and, consequently,

$$\phi = \tfrac{1}{2} Q(Z + \overline{Z}), \tag{3.10}$$

where Q is a constant. Substituting this value of ϕ in $F^{*\mu\gamma}_{j\gamma} n_\mu = 0$, we have verified that this equation is identically satisfied. The electromagnetic energy tensor $E^*_{\mu\gamma}$ is given by

$$E^*_{\mu\gamma} = g^{*\alpha\beta} F^*_{\mu\alpha} F^*_{\gamma\beta} - \tfrac{1}{4} g^*_{\mu\gamma} F^{*\alpha\beta} F^*_{\alpha\beta}. \tag{3.11}$$

Using (3.3) and (3.4) in (3.11) we obtain

$$\begin{aligned} S^4 E^{*J}_\gamma = {} & -\frac{1}{2}\left[\phi^2 \left(\frac{Z - \overline{Z}}{2}\right)^2 - (\phi_{,\rho} l^\rho)^2\right] \delta^J_\gamma \\ & + \phi^2 \left(\frac{Z - \overline{Z}}{2}\right)(m^J \overline{m}_\gamma + \overline{m}^J m_\gamma) \\ & + \tfrac{1}{2}\phi\phi_{,\rho} \overline{m}^\rho (\overline{Z} - Z)(l^J m_\gamma + m^J l_\gamma) \\ & - \tfrac{1}{2}\phi\phi_{,\rho} m^\rho (\overline{Z} - Z)(l^J \overline{m}_\gamma + \overline{m}^J l_\gamma) \\ & + \phi_{,\rho} \phi^{,\rho} l^J l_\gamma - \phi_{,\rho} l^\rho (\phi^{,J} l_\gamma + \phi_{,\gamma} l^J). \end{aligned} \tag{3.12}$$

IV. THE TWO-FLUID INTERPRETATION

In this section we shall try to solve the Einstein–Maxwell equations corresponding to a mixture of perfect fluid, null fluid, and source-free electromagnetic fields in terms of the metric tensor $g^*_{\mu\gamma}$. The energy momentum tensor for such a mixture is given by

$$T^{*J}_{\gamma}=(w+p)V^{*J}V^{*}_{\gamma}+p\delta^{J}_{\gamma}+\Sigma l^{*J}l^{*}_{\gamma}+E^{*J}_{\gamma}, \quad (4.1)$$

where $V^{*\mu}$ are the components of the flow vector satisfying $V^{*\mu}\ V^{*}_{\mu}=-1$. Here w and p are the density and the pressure of the perfect fluid, respectively. Also Σ is the radiation density of the null fluid. Since

$$\delta^{J}_{\gamma}=-l^{J}n_{\gamma}-n^{J}l_{\gamma}+\overline{m}^{J}m_{\gamma}+m^{J}\overline{m}_{\gamma}, \quad (4.2)$$

the Einstein tensor given by (2.13) may be written as

$$\begin{aligned}G^{*J}_{\gamma}&=X(g^{\mu J}-2Hl^{\mu}l^{J})U_{\mu}U_{\gamma}+N^{*}l^{J}l_{\gamma}\\&+[\tfrac{1}{2}X(1+H)-P^{*}+(1/S^{4})\Phi_{0\rho}l^{\rho}]\delta^{J}_{\gamma}\\&+\pi^{*}(l^{J}n_{\gamma}+n^{J}l_{\gamma})-\Delta(\overline{m}^{J}l_{\gamma}+\overline{m}_{\gamma}l^{J})\\&-\overline{\Delta}(m^{J}l_{\gamma}+m_{\gamma}l^{J}), \quad (4.3)\end{aligned}$$

where $X, N^{*}, P^{*}, \Phi_{\theta\mu}, N^{*}$, and N are defined by (2.14). Let us choose the components V^{*}_{μ} of the flow vector as

$$V^{*}_{\mu}=S(\alpha^{*}U_{\mu}+\sqrt{2}\beta^{*}l_{\mu}), \quad (4.4)$$

where

$$\alpha^{*2}=(1+H)/[(1+H)^{2}-(2/X)\{N^{*}-H_{0}S^{-4}\}]^{1/2} \quad (4.5)$$

and

$$\alpha^{*2}(1+H)+2\alpha^{*}\beta^{*}=1. \quad (4.6)$$

One can easily check that

$$g^{*\mu\gamma}V^{*}_{\mu}V^{*}_{\gamma}=V^{*\mu}V^{*}_{\mu}=-1 \quad (4.7)$$

and

$$SV^{*\mu}=\alpha^{*}U^{\mu}+\sqrt{2}l^{\mu}(\beta^{*}+\alpha^{*}H). \quad (4.8)$$

In view of (4.7), the choice (4.4) of V^{*}_{μ} is justified. The expression (4.3) for G^{*J}_{γ} can now be simplified to

$$\begin{aligned}G^{*J}_{\gamma}&=(X/\alpha^{*}2)\ V^{*J}V^{*}_{\gamma}\\&-[\tfrac{1}{2}X(1+H)-P^{*}+(1/S^{4})\Phi_{0\rho}l^{\rho}]\delta^{J}_{\gamma}\\&-\left[N^{*}+\frac{2X\beta^{*}}{\alpha^{*}}\left(1+H+\frac{\beta^{*}}{\alpha^{*}}\right)\right]l^{J}l_{\gamma}\\&+[\pi^{*}-X\beta^{*}/\alpha^{*}](n^{J}l_{\gamma}+l^{J}n_{\gamma})\\&-\Delta(\overline{m}^{J}l_{\gamma}+l^{J}\overline{m}_{\gamma})-\overline{\Delta}(m^{J}l_{\gamma}+l^{J}m_{\gamma}). \quad (4.9)\end{aligned}$$

Now, the Einstein–Maxwell equations are

$$G^{J}_{\gamma}=8\pi T^{*J}_{\gamma}, \quad (4.10)$$

where T^{*J}_{γ} is given by (4.1). Substituting G^{*J}_{γ} from (4.3) and E^{*J}_{γ} from (3.12) in (4.10) one can verify that if

$$X\beta^{*}/\alpha^{*}=\pi^{*}+(8\pi/S^{4})[(\phi_{,\rho}l^{\rho})^{2}+\phi^{2}\Omega^{2}] \quad (4.11)$$

holds, then we have

$$-8\pi(p+w)=X/\alpha^{*2}, \quad (4.12)$$

$$\begin{aligned}8\pi p&=\pi^{*}-\frac{X\beta^{*}}{\alpha^{*}}-\left[\frac{1}{2}X(1+H)-P^{*}+\frac{\Phi_{0\rho}l^{\rho}}{S^{4}}\right]\\&+(4\pi/S^{4})[\phi^{2}\Omega^{2}+(\phi_{,\rho}l^{\rho})^{2}], \quad (4.13)\end{aligned}$$

$$\begin{aligned}-\frac{8\pi\Sigma}{S^{2}}&=-N^{*}-\frac{2X\beta^{*}}{\alpha^{*}}\left(1+H\frac{\beta^{*}}{\alpha^{*}}\right)+\frac{8\pi}{S^{4}}\phi_{,\rho}\phi^{,\rho}\\&+(16/S^{4})\ \pi\phi_{,\rho}l^{\rho}\phi_{,\mu}n^{\mu}, \quad (4.14)\end{aligned}$$

and

$$\Delta+(1/S^{4})[8\pi\phi_{,\rho}l^{\rho}\phi_{,\mu}m^{\mu}+4\pi\phi\phi_{,\mu}m^{\mu}(\overline{Z}-Z)]=0. \quad (4.15)$$

Equation (4.15) can be easily integrated to have the function H_{0} as

$$H_{0}=\tfrac{1}{2}M(Z+\overline{Z})-2\pi Q^{2}Z\overline{Z}+4\pi Q^{2}/R^{2}_{0}, \quad (4.16)$$

where M is a constant. Thus if (4.11) holds we get a solution of the field equations (4.10) with p, w, and Σ given by (4.12), (4.13), and (4.14). The explicit expressions for these physical parameters are

$$\begin{aligned}-8\pi p&=(1/S^{2})(2SS^{11}+S^{12}+1/R^{2}_{0})\\&+(H_{0}/S^{4})(SS^{11}-S^{12}-1/R^{2}_{0}), \quad (4.17)\end{aligned}$$

$$\begin{aligned}-8\pi w&=-(3/S^{2})(S^{12}+1/R^{2}_{0})\\&+(H_{0}/S^{4})(SS^{11}-S^{12}+3/R^{2}_{0})\\&-(2\sqrt{2}/S^{4})S^{1}(H_{0}l^{\rho})_{|\rho}, \quad (4.18)\end{aligned}$$

$$\begin{aligned}-8\pi\Sigma S^{2}&=4H_{0}/R^{2}_{0}+2\sqrt{2}S^{1}(H_{0}n^{\rho})_{|\rho}\\&+2A[-2H/R^{2}_{0}+H_{0}SS^{11}\\&+\sqrt{2}S^{1}(H_{0}l^{\rho})_{|\rho}], \quad (4.19)\end{aligned}$$

where

$$A=1+\frac{H_{0}}{S^{2}}\frac{\{-2H_{0}/R^{2}_{0}+H_{0}SS^{11}+\sqrt{2}S^{1}(H_{0}l^{\rho})_{|\rho}\}}{2S^{2}(SS^{11}-S^{12}-1/R^{2}_{0})} \quad (4.20)$$

and

$$\begin{aligned}(H_{0}l^{\rho})_{|\rho}&=-(H_{0}n^{\rho})_{|\rho}=\tfrac{1}{2}M(Z\overline{Z}-2/R^{2}_{0})\\&+4\pi Q^{2}(Z+\overline{Z})/R^{2}_{0}. \quad (4.21)\end{aligned}$$

The quantity Z is given by (2.9). The explicit form of the line element is given by

$$\begin{aligned}ds^{2}&=S^{2}[-dt^{2}+dr^{2}-2k\sin^{2}\alpha\,d\beta\,dr+(|\rho|^{2}/N^{2})d\alpha^{2}\\&+(|\rho|^{2}+k^{2}\sin^{2}\alpha)\sin^{2}\alpha\,d\beta^{2}]\\&+H_{0}(-dt-dr+k\sin^{2}\alpha\,d\beta)^{2} \quad (4.22)\end{aligned}$$

with

$$\begin{aligned}H_{0}&=-\frac{\sqrt{2}M(R^{2}_{0}-k^{2})}{R_{0}|\rho|^{2}}\sin\left(\frac{r}{R_{0}}\right)\cos\left(\frac{r}{R_{0}}\right)\\&+\frac{4\pi Q^{2}}{R^{2}_{0}}-\frac{4\pi Q^{2}}{|\rho|^{4}}\Big[k^{2}N^{2}\cos^{2}\alpha\\&+\frac{(R^{2}_{0}-k^{2})}{R^{2}_{0}}\sin^{2}\left(\frac{r}{R_{0}}\right)\cos^{2}\left(\frac{r}{R_{0}}\right)\Big]. \quad (4.23)\end{aligned}$$

When $R\to\infty$ and $S=1$, it can be easily seen that Eq. (4.11) is satisfied. In this case we have $p=0$, $w=0$, and $\Sigma=0$. This corresponds to usual Kerr–Newman metric. Thus we can interpret the metric (4.22) as the Kerr–Newman metric in the background of the closed Robertson–Walker universe.

We know that the cosmological constant is nonzero for the Einstein static universe. If we incorporate the cosmological constant Λ in the field equations (4.10), and if we take $S=1$, we have

$$\begin{aligned} -8\pi p &= -\Lambda + (1-H_0)/R_0^2, \\ -8\pi w &= \Lambda - 3(1-H_0)/R_0^2, \quad \Sigma = 0, \end{aligned} \tag{4.24}$$

where H_0 is given by (4.23). In this case Eq. (4.11) is also satisfied.

Thus the case $S=1$ represents the Kerr–Newman metric in the background of the Einstein universe discussed by Patel and Trivedi.[4]

V. THE KERR-NEWMAN METRIC IN EINSTEIN-DE SITTER BACKGROUND

We now consider a special case in which the Robertson–Walker metric has flat spatial sections $t=$ const. Thus we take $R_0 \to \infty$. We shall further assume that the background space-time is pressure-free. Thus we assume that the function S satisfies the differential equation

$$2S^{11}S + S^{12} = 0. \tag{5.1}$$

The solution of (5.1) can be expressed as

$$S = [3(t-t_0)/T_0]^{2/3}, \tag{5.2}$$

where t_0 and T_0 are constants of integration. Let us define T by

$$T = T_0 S^{1/2} = T_0[3(t-t_0)/T_0]^{1/3}. \tag{5.3}$$

It is painless to see that

$$S^1 = 2/T. \tag{5.4}$$

Using $R_0 \to \infty$. In (4.23) we obtain

$$H_0 = (2mr - 4\pi e^2)/|\rho|^2, \tag{5.5}$$

with

$$|\rho|^2 = r^2 + k^2\cos^2\alpha. \tag{5.6}$$

The constants m and e are defined by

$$2m = -\sqrt{2}M \quad \text{and} \quad e = Q. \tag{5.7}$$

In this case, we can easily establish the following results:

$$\begin{aligned} X &= -12/S^2T^2, \quad P^* = -6/S^2T^2, \\ \frac{\beta^*}{\alpha^*} &= -\frac{m(T+r)}{3|\rho|^2S^2} + \frac{2\pi e^2}{3|\rho|^2S^2}, \\ -N^* &= \frac{8m}{TS^4|\rho|^2} - \frac{16\pi e^2k^2\sin^2\alpha}{S^4|\rho|^6}, \\ \Phi_{0\rho}l^\rho &+ \Gamma_0 = 0, \\ E &= -\frac{4m(T+r)}{T^2|\rho|^2} + \frac{8\pi e^2}{T^2|\rho|^2}. \end{aligned} \tag{5.8}$$

Using the results (5.8) it is easy to see that Eq. (4.11) is satisfied. It is a consequence of the above equations (4.12)–(4.14) that

$$8\pi p = 12(mr - 2\pi e^2)/S^4|\rho|^2T^2, \tag{5.9}$$

$$8\pi\omega = \frac{12}{S^2T^2}\left[1 + \frac{m(r-2T)}{3|\rho|^2S^4T^2}\right] - \frac{8\pi e^2}{|\rho|^2S^4T^2}, \tag{5.10}$$

and

$$\begin{aligned} 8\pi\Sigma = \frac{8m}{T^2|\rho|^2S^2}&\left[r + \frac{m(T+r)(5r-T)}{3|\rho|^2S^2}\right] \\ &+ 160\pi^2e^4/3|\rho|^4S^4T^2. \end{aligned} \tag{5.11}$$

The explicit form of the metric in this case is

$$\begin{aligned} ds^2 = S^2(t)[&-dt^2 + dr^2 - 2k\sin^2\alpha\, d\beta\, dr \\ &+ (r^2 + k^2\cos^2\alpha)d\alpha^2 + (r^2+k^2)\sin^2\alpha\, d\beta^2] \\ &+ \left(\frac{2mr - 4\pi e^2}{r^2 + k^2\cos^2\alpha}\right)(-dt - dr + k\sin^2\alpha\, d\beta)^2, \end{aligned} \tag{5.12}$$

where S is given by (5.2).

The Metric (5.12) describes the Kerr–Newman metric in the background of the Einstein–de Sitter universe.

ACKNOWLEDGMENT

The authors are indebted to the referee for his helpful suggestions.

[1] P. C. Vaidya, Pramana **8**, 512 (1977).

[2] R. P. Kerr, Phys. Rev. Lett. **11**, 237 (1963).

[3] A. H. Taub, Ann. Phys. (NY) **134**, 326 (1981).

[4] L. K. Patel and H. B. Trivedi, J. Astrophys. Astro. **3**, 63 (1982).

[5] E. Newman, E. Couch, K. Chinnapared, A. Exton, A. Prakash, and R. Torrence, J. Math. Phys. **6**, 918 (1965).

Structure of the space of solutions of Einstein's equations: A new variables approach

R. V. Saraykar
Department of Mathematics, Nagpur University Campus, Nagpur-440 010, India

(Received 16 June 1987; accepted for publication 31 August 1987)

Following Ashtekar's new Hamiltonian formulation of general relativity ["A new Hamiltonian formulation of general relativity," Syracuse University preprint, 1986; Phys. Rev. Lett. **57**, 2244 (1986); Proceedings of the Florence Conference on "*Constrained Systems*" (World Scientific, Singapore, 1986)], the results of Arms, Fischer, Marsden, and Moncrief [Ann. Inst. H. Poincaré **33**, 147 (1980); Ann. Phys. (NY) **144**, 81 (1982)] on the structure of the space of solutions of vacuum Einstein equations in the case of space-times admitting a compact Cauchy hypersurface are rederived and extended.

I. INTRODUCTION

Recently Ashtekar[1] gave a new Hamiltonian formulation of general relativity using what he called "new variables." One of the variables is a densitized soldering form $\tilde{\sigma}^a{}_A{}^B$, which is an isomorphism between the space of complex tangent vectors V^a at any point of a three-manifold Σ (a Cauchy hypersurface of a space-time) and SU(2) [or SL(2,C)] spinors $V^A{}_B$ at that point whose components form a 2×2 anti-Hermitian traceless matrix. Another conjugate variable is a connection one-form $A_{aM}{}^N$ with values in the Lie algebra of SU(2) corresponding to a certain connection, now known as the Ashtekar–Sen–Witten (ASW) connection (see Renteln[2] and references therein). Using these as dynamical variables, constraints of Einstein's theory in the ADM formalism simply state that $\tilde{\sigma}^a{}_A{}^B$ satisfies the Gauss-law constraint w.r.t. $A_{aM}{}^N$ and that the curvature form $F_{abA}{}^B$ of $A_{aM}{}^N$ satisfies certain purely algebraic conditions involving $\tilde{\sigma}^a{}_A{}^B$. In particular, the constraints are at worst quadratic in these variables. In the ADM formalism, constraints contain nonpolynomial functions of the three-metric. This simplification occurs because $A_{aM}{}^N$ has information about both the three-metric and its conjugate momentum. In the four-dimensional space-time picture, $A_{aM}{}^N$ turns out to be a potential for the self-dual part of Weyl curvature. One of the striking features of this formulation is the simplification of Einstein's equations in the half-flat case. Here, the equations are remarkably simple and resemble Euler's equations for rigid bodies. This supports the conjecture of exact integrability of the half-flat equation. For other simplifications, comments, and discussions, we refer the reader to Ref. 1.

This Hamiltonian formulation can be derived from a manifestly covariant four-dimensional Lagrangian formulation in which A_a (or A_μ, a four-quantity) appears as a connection one-form with values in the Lie algebra of SL(2,C) and $\tilde{\sigma}^a$ is the conjugate momentum $\partial L/\partial \dot{A}_a$ corresponding to a suitable Lagrangian constructed from the curvature form $F_{\mu\nu}$ corresponding to A_μ and $\Sigma_{\mu\nu}{}^{AB} = i(\gamma_\mu{}^{AA'}\gamma_{\nu A'}{}^B - \gamma_\nu{}^{AA'}\gamma_{\mu A'}{}^B)$. The action is

$$I = \int \Sigma_A{}^B \wedge F_B{}^A.$$

The γ giving $\Sigma_{\mu\nu}$ can be constructed as follows. Given a SL(2,C) principal bundle on space-time ${}^4S = \Sigma\times R$, Σ being a spacelike hypersurface, let V denote the two-dimensional complex vector space of Weyl spinors at each point $x\in{}^4S$. Then $V\otimes\bar{V}$ ($\bar{V}$ being the complex conjugate) is also a vector space, the space of 1–1 spinors. If W denotes cotangent space at x on ${}^4S$, then γ is a section of the associated tensor bundle $V\otimes\bar{V}\otimes W$. Thus γ_μ is a Hermitian matrix valued one-form and A_μ is a connection one-form [on the SL(2,C) bundle on ${}^4S$] with values in the Lie algebra of SL(2,C). For further details, see Samuel.[3]

Following this new variables approach we, in this paper, rederive and extend the results of Fisher *et al.*[4] (hereafter referred to as FMM) and Arms *et al.*[5] (hereafter referred to as AMM) on the structure of the space of solutions of Einstein's equations when a space-time admits a compact Cauchy hypersurface. The word "extend" is used in the sense that Ashtekar's formulation gives complex general relativity. Simplicity of constraint equations in new variables simplify calculations of the FMM program (deriving results of FMM and AMM for vaccum Einstein or for Einstein equations coupled to matter fields is referred to as the FMM program) and many features of the FMM program for coupled Einstein–Yang–Mills fields appear in this setting. More simplifications seem to appear in the treatment of the FMM program for the coupled Einstein–Yang–Mills system via new variables. This happens because the new variables formulation provides a natural embedding of the constraint surface of Einstein phase-space into that of Yang–Mills. This embedding seems to provide new tools to analyze a number of issues in both classical and quantum gravity. For detailed comments we refer the reader to Ref. 1, Jacobson and Smolin,[6] and Renteln and Smolin.[7]

In Sec. II, we give notation and a brief introduction to the new variables approach and write the connection one-form, its curvature form, and other identities that will be used later. Again, for details, see Ref. 1. In Sec. III, we prove the first steps of the FMM program, namely ellipticity of the adjoint operator appearing in the evolution equations and linearized stability of new constraint equations by assuming conditions on the spacelike hypersurface Σ similar to Arms[8] and Fischer and Marsden.[9] We then prove a result that enables us to identify the triplet $(N, N_a, \underset{\sim}{N})$ of the lapse func-

0022-2488/88/010186-11$02.50

tion, shift field, and Lie algebra-valued function with symmetrics that a space-time admits. This follows the procedure used by Arms.[8] In Sec. IV, we give the remaining steps of the FMM program, using the Kuranishi map and the conical structure of singularities. Since FMM and AMM contain detailed expositions of these results, we only outline their procedure and give the necessary calculations in terms of the new variables. In Sec. V, we discuss future work related to these results.

II. NEW VARIABLES—A BRIEF INTRODUCTION

Fix a compact three-manifold Σ. The configuration space $\widetilde{C}$ in the ADM-formalism is the space of all positive definite metrics q_{ab} on Σ. Fix a point q_{ab} in $\widetilde{C}$. A tangent vector at q_{ab} is represented by a second rank symmetric tensor field h_{ab} on Σ. A cotangent vector is therefore represented by a second rank symmetric tensor density p^{ab} of weight 1. The phase space $\widetilde{\Gamma}$ of classical general relativity (ADM) is the cotangent bundle over $\widetilde{C}$. Thus a point of $\widetilde{\Gamma}$ is a pair (q_{ab},p^{ab}). Here $\widetilde{\Gamma}$ has a natural symplectic structure $\widetilde{\Omega}$ whose action at a point (q,p) of $\widetilde{\Gamma}$ on tangent vectors (h,w) and (h',w') at that point is given by

$$\widetilde{\Omega}|_{(q,p)}((h,w),(h',w'))=\int(w\cdot h'-w'\cdot h), \tag{1}$$

where $w\cdot h=w^{ab}h_{ab}$. Not all points of $\widetilde{\Gamma}$ are accessible to the vacuum gravitational field: There are constraints given by

$$C_a(q,p)=-2q_{am}D_np^{mn}=0, \tag{2}$$

$$C(q,p)=-\frac{\sqrt{q}}{G}R+\frac{G}{\sqrt{q}}\left(p^{ab}p_{ab}-\frac{1}{2}p^2\right)=0, \tag{3}$$

where D and R are, respectively, the derivative operator and the scalar curvature of q_{ab}, $\sqrt{q}=\sqrt{\det q}$, $p^2=(\operatorname{tr}p)^2$, and $\operatorname{tr}p=q_{ab}\,p^{ab}$. These constraint equations and corresponding evolution equations which together are equivalent to Einstein field equations, are treated in detail in ADM[10] and Fischer and Marsden.[9] We shall follow Refs. 8 and 9 for notation and other technical details.

Phase space $\widetilde{\Gamma}$ is then extended to incorporate spinor fields. Here Σ is not equipped with an *a priori* metric. So we first spell out the sense in which the fields are spinorial.

In addition to tensor fields $t^{a\cdots b}{}_{c\cdots d}$ on Σ, we also consider objects λ^A,μ_A with internal SU(2) indices. Mathematically these fields are cross sections of a vector bundle over Σ whose fibers are two-(complex) dimensional vector spaces, equipped with preferred nondegenerate two-forms ϵ^{AB} and ϵ_{AB}. We shall raise and lower the internal indices with these forms:

$$\lambda^A=\epsilon^{AB}\lambda_B,\quad \mu_A=\mu^B\epsilon_{BA}. \tag{4}$$

We now introduce soldering forms that tie the abstractly defined internal indices to the tangent space of Σ, thereby making them spinor indices. Consider isomorphisms $\sigma_a{}^{AB}$ from the tangent vectors λ^a to Σ to the trace-free, second rank, Hermitian spinors λ^{AB}: $\lambda^{AB}=\sigma_a{}^{AB}\lambda^a$. Denote the inverse mapping by $\sigma^a{}_{AB}$. Properties of ϵ, the Hermitian conjugation, and σ imply that

$$q_{ab}=\sigma_a{}^{AB}\sigma_b{}^{CD}\epsilon_{AC}\epsilon_{BD}=-\operatorname{Tr}\sigma_a\sigma_B \tag{5}$$

is a positive definite three-metric on Σ. Thus, given a specific σ, we can go back to the standard spinorial analysis (see the Appendix of Ref. 1). This σ (densitized) is the basic dynamical variable of Ashtekar's approach. Metric q_{ab} is to be thought of as a secondary object, derived from the primary dynamical variable $\sigma_a{}^{AB}$.

The new extended phase space Γ can now be defined. Fix a soldering form $\sigma^a{}_{AB}$ (and its inverse $\sigma_a{}^{AB}$) whose connection D is flat. Let C be the space of all soldering forms $\sigma^a{}_{AB}$. Thus C is the new configuration space. Given any σ in C, we obtain a q_{ab} in $\widetilde{C}$ via (5). If σ_1 and σ_2 project down the same metric q_{ab} then they are related by a local SU(2) transformation. Thus the enlargement of the configuration space from C to $\widetilde{C}$ is brought about because of the freedom to perform internal SU(2) rotations. While q_{ab} has six real components per space point, $\sigma^a{}_{AB}$ has nine; the new three degrees of freedom correspond to precisely the three SU(2) rotations.

The momentum conjugate to $\sigma^a{}_{AB}$ is a density of weight 1, $M_a{}^{AB}$, whose index structure is opposite to that of $\sigma^a{}_{AB}$. The action of the cotangent vector $M_a{}^{AB}$ on any tangent vector $(\delta\sigma)^a{}_{AB}$ at a point $\sigma^a{}_{AB}$ of C is given by

$$M(\delta\sigma)=\int_\Sigma M_a{}^{AB}(\delta\sigma)^a{}_{AB}. \tag{6}$$

The extended phase space Γ is the cotangent bundle over C. Thus a point of Γ is a pair $(\sigma^a{}_{AB},M_a{}^{AB})$. The natural symplectic structure Ω on Γ is given by

$$\Omega_{(\sigma,M)}((\delta\sigma,\delta M),(\delta\sigma',\delta M'))$$
$$=\int_\Sigma(\delta M_a{}^{AB})(\delta\sigma'^a{}_{AB})-(\delta M'_a{}^{AB})(\delta\sigma^a{}_{AB}), \tag{7}$$

where $(\delta\sigma,\delta M)$ and $(\delta\sigma',\delta M')$ are any two tangent vectors at the point (σ,M) of Γ. The Hamiltonian vector field and Poisson brackets are given as usual.

We now describe the constraints. From the Hamiltonian viewpoint, the SU(2) rotations are gauge motions, hence their generating functionals should vanish. So the three new constraints are

$$C_{ab}=M_{[a}{}^{AB}\sigma_{b]AB}\equiv M_{[ab]}=0$$

or (8)

$$C^{AB}=\sigma^{aM(A}M_a{}^{B)N}\epsilon_{MN}=0.$$

Set $M^{ab}=p^{ab}$. (In the noncompact case, $M^{(ab)}=p^{ab}$ and appropriate boundary conditions are to be satisfied by p^{ab} in order that this holds.) Then the remaining constraints are the standard ones:

$$C_a(\sigma,M)=-2q_{am}D_n\,p^{mn}=0, \tag{2'}$$

$$C(\sigma,M)=-\sqrt{q}R+(1/\sqrt{q})(p^{ab}p_{ab}-\tfrac{1}{2}p^2)=0. \tag{3'}$$

Hereafter, we take $G=1$ since we are not interested in strong coupling limit.

Thus we now have $3+3+1=7$ components. The configuration variable $\sigma^a{}_{AB}$ has nine components per space point. Thus we have two degrees of freedom per space point. Canonical transformations generated by (2) and (3) retain their usual meaning. Modulo the new constraints (8), q_{ab} and p^{ab} have the same Poisson brackets (see Ref. 1). Thus the enlargement of the phase space $\tilde{\Gamma}$ is compatible with the symplectic structure $\tilde{\Omega}$.

In transition from $\tilde{\Gamma}$ to Γ, the constraint equations (2) and (3) have remained intact. The addition of new degrees of freedom does not by itself simplify the constraints. The key step in the simplification is the introduction of new variables. The extension to Γ is necessary because these variables cannot be defined on $\tilde{\Gamma}$.

Fix a point $(\sigma^a{}_{AB}, M_a{}^{AB})$ of Γ. Introduce two connections ${}^{\pm}\mathscr{D}$, which act on tensor and spinor fields on (Σ,σ):

$$ {}^{\pm}\mathscr{D}_a\lambda_{bM} = D_a\lambda_{bM} \pm (i/\sqrt{2})\Pi_{aM}{}^N\lambda_{bN} \tag{9}$$

where D_a is the connection that annihilates the given $\sigma^a{}_{AB}$ and where $\Pi_{aM}{}^N$ is given by

$$\Pi_{aM}{}^N = q^{-1/2}(M_{aM}{}^N - \tfrac{1}{2}M_b{}^{AB}\sigma^b{}_{AB}\sigma_{aM}{}^N)$$

or

$$M_{aM}{}^N = \sqrt{q}(\Pi_{aM}{}^N - \Pi_b{}^{AB}\sigma^b{}_{AB}\sigma_{aM}{}^N). \tag{10}$$

Thus $\Pi_a{}^{AB}$ is related to $M_a{}^{AB}$ in the same way that the extrinsic curvature k_{ab} is related to p^{ab}:

$$p^{ab} = \sqrt{q}(k^{ab} - k^{mn}q_{mn}q^{ab}). \tag{11}$$

Note that $\Pi_{ab} = -\operatorname{Tr}\Pi_a\sigma_b$ is not necessarily symmetric.

As in gauge theories, it is convenient to work with connection one-forms $A_{aA}{}^B$ in place of derivative operators. So, fix a fiducial connection ∂_a and for simplicity assume that ∂_a commutes with Hermitian conjugation, $\partial_a\lambda^{\dagger}_B = (\partial_a\lambda_B)^{\dagger}$ and has zero internal curvature, $\partial_{[a}\partial_{b]}\lambda_A = 0$. Set

$$ {}^{\pm}\mathscr{D}_a\lambda_{bM} = \partial_a\lambda_{bM} + {}^{\pm}A_{aM}{}^N\lambda_{bN} \tag{12}$$

so that (9) gives

$$ {}^{\pm}A_{aM}{}^N = \Gamma_{aM}{}^N \pm (i/\sqrt{2})\Pi_{aM}{}^N, \tag{13}$$

where Γ are the spin connection one-forms of D: $(D_a - \partial_a)\lambda_M = \Gamma_{aM}{}^N\lambda_N$.

Thus ${}^{\pm}A$ contains information about both σ and M and ${}^{+}A$ or ${}^{-}A$ is one of the new variables. It follows that (see Ref. 1) ${}^{+}A$ (or ${}^{-}A$) constitute a set of commuting variables. Also Poisson brackets between $\sigma^a{}_{AB}$ and $A_a{}^{AB}$ are simple. Set $\tilde{\sigma}^a{}_{AB} = \sqrt{q}\sigma^a{}_{AB}$. Then

$$\{\tilde{\sigma}^a{}_{AB}(x), \tilde{\sigma}^m{}_{MN}(x)\} = 0.$$

Using the fact that $\Pi_a{}^{AB}$ and $\tilde{\sigma}^m{}_{MN}$ are canonically conjugate,

$$\{\Pi_a{}^{AB}(x), \tilde{\sigma}^m{}_{MN}(y)\} = \delta_a{}^m\delta^{(A}{}_M\delta^{B)}{}_N\delta(x,y),$$

it follows that

$$\{{}^{\pm}A_a{}^{AB}(x), \tilde{\sigma}^m{}_{MN}(y)\} = \pm(i/\sqrt{2})\delta^{(A}{}_M\delta^{B)}{}_N\delta(x,y). \tag{14}$$

Thus $\tilde{\sigma}^a$ may be thought of as being "canonically conjugate" to ${}^{\pm}A_a$. $\tilde{\sigma}^a$ and ${}^{\pm}A_a$ are Ashtekar's new dynamical variables.

A. Revised constraint equations

Constraint equations (8), (2), and (3) can now be reexpressed in terms of the new variables $\tilde{\sigma}^a$ and ${}^{\pm}A_a$.

It turns out that

$$ {}^{\pm}\mathscr{D}_a\tilde{\sigma}^a{}_{AB} = \pm\sqrt{2}iM_{[ab]}\sigma^a{}_A{}^M\sigma^b{}_{MB}.$$

Hence (8) is equivalent to

$$ {}^{\pm}\mathscr{D}_a\tilde{\sigma}^a{}_A{}^B = 0. \tag{8$'$}$$

Since the divergence of a tensor density of weight 1 is independent of the choice of the derivative operator, $\mathscr{D}_a\tilde{\sigma}^a{}_A{}^B$ can be expanded knowing only the action

$$ {}^{\pm}\mathscr{D}_a\lambda_M = \partial_a\lambda_M + {}^{\pm}A_{aM}{}^N\lambda_N \tag{15}$$

of ${}^{\pm}\mathscr{D}$ on internal indices. We then have

$$ {}^{\pm}\mathscr{D}_a\tilde{\sigma}^a{}_A{}^B \equiv \partial_a\tilde{\sigma}^a{}_A{}^B + [{}^{\pm}A_a, \tilde{\sigma}^a]_A{}^B = 0. \tag{8$''$}$$

Thus (8) has been reexpressed in terms of the new variables.

Next, define spinorial curvature of ${}^{\pm}\mathscr{D}$ by

$$ {}^{\pm}F_{abM}{}^N\lambda_N = 2\,{}^{\pm}\mathscr{D}_{[a}\mathscr{D}_{b]}\lambda_M, \tag{16}$$

so that

$$ {}^{\pm}F_{abM}{}^N = 2\partial_{[a}\,{}^{\pm}A_{b]M}{}^N + [{}^{\pm}A_a, {}^{\pm}A_b]_M{}^N. \tag{16$'$}$$

Thus using (9), one obtains

$$ {}^{\pm}F_{abc} = R_{abc} - (1/\sqrt{2})\epsilon_{cde}\Pi_a{}^d\Pi_b{}^e \pm \sqrt{2}iD_{[a}\Pi_{b]c}, \tag{17}$$

where R_{abc} is the spinorial curvature of D. We have $R_{ab}{}^{cd} = -\sqrt{2}R_{abp}\epsilon^{pcd}$. Then it follows that

$$\begin{aligned}\operatorname{Tr}\sigma^a\,{}^{\pm}F_{ab} &= (1/2\sqrt{2})(\Pi_{am}\Pi_{bn} - \Pi_{bm}\Pi_{an})\epsilon^{mna}\\ &\quad\mp(i/\sqrt{2})D^a(\Pi_{ba} - \Pi q_{ba})\\ &\simeq \mp(i/\sqrt{2})D^a(k_{ab} - kq_{ab}),\end{aligned} \tag{18}$$

where $\simeq$ denotes the equality modulo constraint (8). Thus constraint (2) can be rewritten as

$$0 = C_a(\tilde{\sigma},A) \equiv -2q_{am}D_n\,p^{mn} = \mp 2\sqrt{2}i\operatorname{Tr}\tilde{\sigma}^mF_{ma} \tag{2$''$}$$

in terms of the new basic variables $\tilde{\sigma}^a{}_{AB}$ and ${}^{\pm}A_a{}^{AB}$. For constraint (3) we note that

$$\begin{aligned}\operatorname{Tr}\sigma^a\sigma^bF_{ab} &= -(1/\sqrt{2})\epsilon^{abc}F_{abc}\\ &= \tfrac{1}{2}(R + \Pi^2 - \Pi_{ab}\Pi^{ba}) \mp i\epsilon^{abc}D_a\Pi_{bc}\\ &\simeq \tfrac{1}{2}(R + k^2 - k_{ab}k^{ab}).\end{aligned} \tag{19}$$

Hence (3) becomes

$$0 = -2q^{-1/2}\operatorname{Tr}\tilde{\sigma}^a\tilde{\sigma}^bF_{ab}. \tag{3$''$}$$

Thus the set of Einstein constraint equations can be rewritten in terms of the new variables as

$$C_G(\tilde{\sigma},A) \equiv -\sqrt{2}i\mathscr{D}_a\tilde{\sigma}^a{}_A{}^B = 0, \tag{8$'''$}$$

$$C_a(\tilde{\sigma},A) \equiv -2\sqrt{2}i\operatorname{Tr}\tilde{\sigma}^mF_{ma} = 0, \tag{2$'''$}$$

$$C(\tilde{\sigma},A) \equiv -2\operatorname{Tr}\tilde{\sigma}^a\tilde{a}^bF_{ab} = 0. \tag{3$'''$}$$

For remarks on these polynomial constraint equations see the first paper cited in Ref. 1.

B. Hamiltonian evolution equations

Evolution equations can be written by the usual procedure using the full Hamiltonian of the theory:

$$\begin{aligned}H = \int_{\Sigma}\operatorname{Tr}A_a\frac{\partial\tilde{\sigma}^a}{\partial t} - NC(\tilde{\sigma},A) - N^aC_a(\tilde{\sigma},A)\\ - \underline{N}_A{}^B(C_G(\tilde{\sigma},A))_B{}^A.\end{aligned} \tag{20}$$

Here N is a lapse function of density minus one, N^a is a shift vector field, and $\underline{N}_A{}^B$ is a function on Σ with values in the Lie algebra of SU(2).

The same Hamiltonian can also be obtained from the manifestly covariant four-Lagrangian density

$$\tilde{L} = \tfrac{1}{2}\tilde{\eta}^{\mu\nu\alpha\beta}\,\mathrm{Tr}(\Sigma_{\mu\nu}F_{\alpha\beta}),$$

$\tilde{\eta}^{\mu\nu\alpha\beta}$ being the Levi-Civita tensor density as given by Samuel.[3]

In any case, the evolution equations can be written in compact notation as in Fischer–Marsden[9]

$$\frac{\partial}{\partial t}\begin{bmatrix}\tilde{\sigma}\\ A\end{bmatrix} = -\,\mathrm{J}\circ D\Phi(\tilde{\sigma},A)^*\begin{bmatrix}N\\ N^a\\ \underline{N}_A{}^B\end{bmatrix}. \tag{21}$$

Here J is the complex structure

$$\begin{bmatrix}0 & -I\\ I & 0\end{bmatrix}$$

and

$$\Phi(\tilde{\sigma},A) = (C(\tilde{\sigma},A),C_a(\tilde{\sigma},A),C_G(\tilde{\sigma},A)) \tag{22}$$

is regarded as a mapping from $(\tilde{\chi}\otimes\mathscr{G},\Lambda^1\otimes\mathscr{G})$ into $(\Lambda^{0*},\chi^*,(\overline{\Lambda}^0\otimes\mathscr{G})^{\sim *})$, where $\mathscr{G}$ = Lie algebra of SU(2) and

$\tilde{\chi}\otimes\mathscr{G}$ = {smooth tensorial $\mathscr{G}$-valued vector densities on Σ},
$\Lambda^k\otimes\mathscr{G}$ = {smooth tensorial $\mathscr{G}$-valued k-forms on Σ},
Λ^{0*} = {smooth scalar densities of weight 2 on Σ},
χ^* = {smooth one-form densities on Σ},
$\overline{\Lambda}^0\otimes\mathscr{G}$ = {smooth tensorial $\mathscr{G}$-valued scalar densities on Σ},
$(\overline{\Lambda}^0\otimes\mathscr{G})^{\sim}$ = quotient of $\overline{\Lambda}^0\otimes\mathscr{G}$ by constant function densities with values in the center of $\mathscr{G}$,
$(\overline{\Lambda}^0\otimes\mathscr{G})^{\sim *}$ = {images of $\tilde{\sigma}$ under $\mathscr{D}_a$}

(see Ref. 8, p. 446 for more explanation), $D\Phi$ denotes the Fréchet derivative of Φ and $D\Phi^*$ is the L^2-adjoint of the linear operator $D\Phi(\tilde{\sigma},A)$. For more details about this form of evolution equations, see Ref. 9.

Explicitly, evolution equations are given as

$$\frac{\partial\tilde{\sigma}^b}{\partial t} = 4\mathscr{D}_a(N\tilde{\sigma}^{[a}\tilde{\sigma}^{b]}) + 4\sqrt{2}i\mathscr{D}_a(N^{[a}\tilde{\sigma}^{b]}) - \sqrt{2}i[\underline{N},\tilde{\sigma}^b], \tag{23}$$

$$\frac{\partial A_a}{\partial t} = 2[N\tilde{\sigma}^b,F_{ab}] - 2\sqrt{2}iN^bF_{ab} - \sqrt{2}i\mathscr{D}_a\underline{N}. \tag{24}$$

In the functional derivative notations,

$$D\Phi(\tilde{\sigma},A)^*\cdot(N,N^a,\underline{N}) = \begin{bmatrix}\dfrac{\delta}{\delta\sigma}(NC + N^aC_a + \mathrm{Tr}\,\underline{N}C_G)\\[2ex] \dfrac{\delta}{\delta A}(NC + N^aC_a + \mathrm{Tr}\,\underline{N}C_G)\end{bmatrix}, \tag{25}$$

so that

$$\frac{\partial\tilde{\sigma}}{\partial t} = \frac{\delta}{\delta A}(NC + N^aC_a + \mathrm{Tr}\,\underline{N}C_G) \tag{23'}$$

and

$$\frac{\partial A}{\partial t} = -\frac{\delta}{\delta\sigma}(NC + N^aC_a + \mathrm{Tr}\,\underline{N}C_G). \tag{24'}$$

Here we follow Ashtekar's formulation where the lapse naturally arises as a density of weight minus one in calculating the functional derivatives (i.e., adjoints) because the scalar constraint is a density of weight 2. Thus the integration can be carried out without reference to a specific volume element.

III. LINEARIZED THEORY

A. Ellipticity of $D\Phi^*$ and linearized stability

The main result of this section is the following.

Theorem 1: Constraint equations (8'), (2'), and (3') are linearization stable at $(\tilde{\sigma},A)$ if the following conditions are satisfied on Σ: (1) $q_{ab}\Pi^{ab}$ is constant, (2) Π^{ab} is not identically zero or q_{ab} is not flat (i.e., σ does not correspond to a flat q_{ab}); (3) if N^a is a vector field on Σ and $\underline{N}_A{}^B$ is a $\mathscr{G}$ valued function on Σ such that $\mathscr{L}_N\tilde{\sigma}^b - \frac{1}{4}[\underline{N},\tilde{\sigma}^{\bar{b}}] = 0$ and $N^bF_{ab} = \frac{1}{2}\mathscr{D}_a\underline{N}$, then $N^a = 0$ and the image of $\underline{N}$ lies in the center of $\mathscr{G}$.

Our Π^{ab} here coincides with the Fischer–Marsden[9] k^{ab}; $\mathscr{L}_N\tilde{\sigma}$ is the gauge covariant Lie derivative of $\tilde{\sigma}$, and $\mathscr{L}_NA \equiv N^aF_{ab}$ is the gauge covariant Lie derivative of A (cf. Ref. 8).

As we shall see below, the linearization stability of constraint equations is equivalent to that of Einstein field equations. For a definition of linearization stability and other theoretical details, see Ref. 9 and FMM. Before we give the proof of Theorem 1, we derive some useful facts regarding linearized evolution equations that will be needed later. We work with the space-time metric

$${}^4g_{\mu\nu}\,dx^\mu\,dx^\nu = -(N^2 - N^aN_a)dt^2 + 2N_a\,dx^a\,dt + q_{ab}\,dx^a\,dx^b, \tag{26}$$

where $q_{ab} = -\mathrm{Tr}\,\sigma_a\sigma_b$. Let ${}^4h_{\mu\nu}$ be a solution of the linearized Einstein equations. Let A_μ denote connection one-form on the pricipal SU(2) [or SL(2,C)] bundle on space-time ${}^4S$ (cf. Ref. 3) with values in the Lie algebra of SU(2) or SL(2,C). [For reduction of SL(2,C) spinors to SU(2) spinors see Ref. 1 or Renteln[2] and Sen.[11] So we start with ${}^4g$ and ${}^4h$. The variables $q, p, N, N^{a\cdot}\ \underline{N}$ are uniquely and differentiably defined by the variables ${}^4g$ and A_μ and their derivatives, and vice versa (cf. Arms,[8] Sec. 3). Thus the linearized theorem for Lagrangian variables (see below) is equivalent to a result for the linearized evolution (Hamiltonian) equations,

$$\frac{\partial}{\partial t}\begin{bmatrix}\eta\\ B\end{bmatrix} = -\,\mathrm{J}\circ D\left[D\Phi(\tilde{\sigma},A)^*\begin{bmatrix}N\\ N^a\\ \underline{N}\end{bmatrix}\right]\cdot\begin{bmatrix}\eta\\ B\end{bmatrix} - \mathrm{J}\circ D\Phi(\tilde{\sigma},A)^*\begin{bmatrix}L\\ Y\\ V\end{bmatrix}, \tag{27}$$

where (η,B) are linearized $(\tilde{\sigma},A)$ and (L,Y,V) are linearized $(N,N^a,\underline{N})$. Equation (27) follows similarly as in Ref. 9, Sec. 4. For initial data for the linearized theorem, we have

(η,B) on Σ satisfying the linearized constraint equation

$$D\Phi(\tilde{\sigma},A)\cdot(\eta,B)=0. \tag{28}$$

To construct initial data for the Lagrangian version, we need, in addition, to specify (L,Y,V) and their first derivatives on Σ. From these, we define ${}^4h$ on Σ by

$$\begin{aligned}&{}^4h_{ab}=h_{ab},\quad {}^4h_{0a}=-(\eta_{ba}+\eta_{ab})N^b-Y_a,\\&{}^4h_{00}=-2NL+2N_aY^a+h_{ab}N^aN^b.\end{aligned} \tag{29}$$

Here $\eta_{ab}=-\mathrm{Tr}\,\eta_a\sigma_b$. With the help of ${}^4h$, we find $w=$ linearized p given by $DP_i({}^4g)\cdot{}^4h=(h,w)$, where $P_i({}^4g)=(q,p)$. Then $B_{ab}=-\mathrm{Tr}\,B_a\sigma_b$ is given by

$$B_{ab}=\epsilon_a{}^{mn}\,\partial_m h_{nb}+iw_{ab}.$$

From this, $B_{aM}{}^N$ can be determined. With this $B_{aM}{}^N$ we can now define $B_{\mu M}{}^N$ (linearized A_μ) by

$${}^4B_{aM}{}^N=B_{aM}{}^N,\quad {}^4B_{0M}{}^N=-V_M{}^N. \tag{29'}$$

Thus the background metric ${}^4g$ and A_μ determine $(N,N^a,\underline{N})$ everywhere $(A_{0M}{}^N=-\underline{N}_M{}^N)$. Using (27), we can find $\partial h_{ab}/\partial t$ and $\partial B_{aM}{}^N/\partial t$ on Σ. So, we know the first derivatives of h, B, N, N^a, L, Y, V on Σ, and by differentiating (29) and (29'), we can determine $\partial\,{}^4h_{\mu\nu}/\partial t$ or $\partial\,{}^4B_{\mu M}{}^N/\partial t$. Thus we get the following result analogous to Proposition 2B of Ref. 8.

Lemma 1: Suppose ${}^4g$ is a solution of the Einstein equation on ${}^4S$ and we have initial data on Σ for the linearized system as above. Then there is a solution ${}^4h$ that determines solution ${}^4B$ on ${}^4S$ of the linearized Lagrangian field equations with the given initial data. Any two such ${}^4h$ differ by a linearized coordinate transformation and any two such ${}^4B$ differ by a linearized coordinate and gauge transformation:

$$\begin{aligned}&{}^4\tilde{h}-{}^4h=L_{{}^4\tilde{Y}}{}^4g,\\&{}^4\tilde{B}-{}^4B=\mathscr{L}_{{}^4\tilde{Y}}{}^4A-D\,{}^4\tilde{V},\end{aligned} \tag{30}$$

with ${}^4\tilde{Y}_\mu$, $(\partial/\partial t)\,\tilde{Y}_\mu$, ${}^4\tilde{V}$, and $(\partial/\partial t)\,{}^4\tilde{V}$ are all zero on Σ. Equivalently, for each choice of gauge $(N,N^a,\underline{N})$ and linearized gauge (L,Y,V) on ${}^4S=\Sigma\times R$, the linearized equations (27) have a unique solution (η,B) [or (h,w)].

Remark: In contrast to the Einstein–Yang–Mills system treated by Arms,[8] quantities ${}^4h$ and ${}^4B$ appearing here are not independent but are related through their projected three-dimensional quantities B_{ab}, w_{ab}, and h_{ab}.

Proof of Lemma 1 proceeds exactly as in Ref. 8, Proposition 2B, pp. 448 and 449, except that ${}^4h$ and ${}^4B$ are interrelated. First ${}^4\tilde{Y}$ is determined uniquely through a certain hyperbolic equation and then using ${}^4\tilde{Y}$, ${}^4\tilde{V}$ is determined uniquely by another hyperbolic equation. They satisfy Eqs. (30). Thus we can conclude (Refs. 8 and 9) as follows.

Lemma 2: Einstein field equations are linearization stable at ${}^4g$ if and only if the constraint equations are linearization stable at $(\tilde{\sigma},A)$.

We now proceed to give the proof of Theorem 1. Again, as in Refs. 8 and 9, it is sufficient to show that $D\Phi(\tilde{\sigma},A)^*$ is an elliptic operator and then to show that under the conditions of Theorem 1, $D\Phi(\tilde{\sigma},A)^*$ is injective. This will give the linearization stability of the constraint equations $\Phi(\tilde{\sigma},A)=0$ and hence by Lemma 2, stability of the Einstein field equations at any solution ${}^4g$ that projects the data $(\tilde{\sigma},A)$ on the hypersurface through (q,p). Recall that A is determined by extrinsic curvature and spin connection through Eq. (13).

$D\Phi(\tilde{\sigma},A)^*$ is elliptic:

$$D\Phi(\tilde{\sigma},A)^*\cdot\begin{bmatrix}N\\N^a\\\underline{N}\end{bmatrix}=\begin{bmatrix}-2[N\tilde{\sigma}^b,F_{ab}]+2\sqrt{2}iN^bF_{ab}+\sqrt{2}i\mathscr{D}_a\underline{N}\\4\mathscr{D}_\alpha(N\tilde{\sigma}^{[a}\tilde{\sigma}^{b]})+4\sqrt{2}i\mathscr{D}_\alpha(N^{[a}\tilde{\sigma}^{b]})-\sqrt{2}i[\underline{N},\tilde{\sigma}^b]\end{bmatrix}. \tag{31}$$

The principal symbol of $D\Phi(\tilde{\sigma},A)^*$, for each vector $\xi\in T^*\Sigma$, is given by the map

$$\begin{aligned}S(\xi)\colon&(N,N^a,\underline{N}_A{}^B)\\&\mapsto(N\xi_a\tilde{\sigma}^{[a}\tilde{\sigma}^{b]}-i\tilde{\sigma}^a\xi_aN^b+i\xi_aN^a\tilde{\sigma}^b,i\xi_a\underline{N}_A{}^B).\end{aligned}$$

Here, $D\Phi(\tilde{\sigma},A)^*$ will be elliptic if we show that the map $S(\xi)$ is injective for every $\xi\neq0$. So, if $\xi\neq0$ and $S(\xi)(N,N^a,\underline{N}_A{}^B)=0$, then we have

$$N\xi_a(\tilde{\sigma}^{[a}\tilde{\sigma}^{b]})_A{}^B-i\tilde{\sigma}^a{}_A{}^B\xi_aN^b+i\xi_aN^a\tilde{\sigma}^b{}_A{}^B=0$$

and $\xi_a\underline{N}_A{}^B=0$. Since $\xi\neq0$, we get from the second equation,

$$\xi^a\xi_a\underline{N}_A{}^B=0\ \text{ or }\ \|\xi\|^2\underline{N}_A{}^B=0,\ \text{ or }\ \underline{N}=0. \tag{32a}$$

Multiplying the first equation by $\tilde{\sigma}^c$ and taking Tr, we get

$$-\sqrt{2}N\xi_a\epsilon^{abc}+i(q^{ac}\xi_aN^b-q^{bc}\xi_aN^a)=0.$$

Equating the real part to zero, we get $N\xi_a\epsilon^{abc}=0$, which gives $N\xi_a\epsilon^{abc}\epsilon_{bcp}=0$ or $N\xi_p=0$ and so, as before, since $\xi\neq0$, we get

$$N=0. \tag{32b}$$

Finally, equating the imaginary part to zero, we get $\xi^cN^b-q^{bc}\xi_aN^a=0$, which, after contracting by q_{bc}, gives $-2\xi_bN^b=0$ or $\xi_bN^b=0$. So substituting in $\xi^cN^b-q^{bc}\xi_aN^a=0$, we get $\xi^cN^b=0$, or $\xi_c\xi^cN^b=0$, or $\|\xi\|^2N^b=0$. So, again, since $\xi\neq0$ we get

$$N^b=0. \tag{32c}$$

Thus from (32a)–(32c) we get

$$S(\xi)(N,N^a,\underline{N}_A{}^B)=0\Rightarrow N=0,\quad N^a=0,\quad \underline{N}_A{}^B=0.$$

Hence $D\Phi(\tilde{\sigma},A)^*$ is elliptic.

$D\Phi(\tilde{\sigma},A)^*$ is injective under conditions of Theorem 1: Assume $D\Phi(\tilde{\sigma},A)^*\cdot(N,N^a,\underline{N})=0$. This gives, from Eq. (31),

$$4\mathscr{D}_a(N\tilde{\sigma}^{[a}\tilde{\sigma}^{b]})+4\sqrt{2}i\mathscr{D}_a(N^{[a}\tilde{\sigma}^{b]})-\sqrt{2}i[\underline{N},\tilde{\sigma}^b]=0 \tag{31'}$$

and

$$2[N\tilde{\sigma}^b,F_{ab}]-2\sqrt{2}iN^bF_{ab}+\sqrt{2}i\mathscr{D}_a\underline{N}=0. \tag{31''}$$

Multiplying (31') by $\tilde{\sigma}^d$, taking Tr, using the expression

$$\begin{aligned}\mathscr{D}_a\sigma^b{}_{AB}=D_a\sigma^b{}_{AB}&+(i/\sqrt{2})\Pi_{aA}{}^M\sigma^b{}_{MB}\\&+(i/\sqrt{2})\Pi_{aB}{}^M\sigma^b{}_{AM},\end{aligned}$$

and keeping in mind that D_a annihilates σ^b, we get

$$\begin{aligned}&-\sqrt{2}D_aN\epsilon^{dab}+Ni(\Pi^{bd}-(\operatorname{tr}\Pi)q^{bd})\\&\quad+i(-q^{db}D_aN^a+D^dN^b-(i/\sqrt{2})N^a\Pi_{ac}\epsilon^{cb}{}_pq^{pd})\\&\quad+\sqrt{2}i\underline{N}_c\epsilon^{bdc}=0.\end{aligned}$$

Here we have used the constraint equation $\mathscr{D}_a\tilde{\sigma}^a=0$ and the fact that Π_{ab} becomes an extrinsic curvature if $\mathscr{D}_a\tilde{\sigma}^a=0$; and also the identity $\epsilon^{abc}=-\sqrt{2}\operatorname{Tr}\sigma^a\sigma^b\sigma^c$. Finally, $\underline{N}_c=+\underline{N}_A{}^B\sigma_{cB}{}^A$. The above equation can be written as

$$\begin{aligned}&-\sqrt{2}((D_aN)\epsilon^{dab}-\tfrac{1}{2}N^a\Pi_{ac}\epsilon^{cbd})\\&\quad+i(N(\Pi^{bd}-(\operatorname{tr}\Pi)q^{bd})+D^dN^b-(D_aN^a)q^{bd})\\&\quad-\sqrt{2}i\underline{N}_c\epsilon^{bdc}=0.\end{aligned}\tag{31'''}$$

Contracting this equation with q^{bd}, we get

$$D_aN^a=-N(\operatorname{tr}\Pi).\tag{33}$$

Equating the imaginary part of (31‴) to zero and using (33) we get

$$D^dN^b=-N\Pi^{bd}+\sqrt{2}\underline{N}_c\epsilon^{bdc}.\tag{34}$$

Now, if we equate the real part of (31‴) to zero, and contract by ϵ_{dpb}, we get

$$D_pN-\tfrac{1}{2}N^a\Pi_{ap}=0.$$

Operating on D^p gives

$$\Delta N-\tfrac{1}{2}D^p(N^a\Pi_{ap})=0$$

or

$$\Delta N-\tfrac{1}{2}(D^pN^a)\Pi_{ap}-\tfrac{1}{2}N^aD^p\Pi_{ap}=0.$$

By the constraint equation $C_a=0$, $D^p\Pi_{ap}=D^p(\operatorname{tr}\Pi)q_{ap}$, and by our assumption, $\operatorname{tr}\Pi=\text{const}$, this becomes zero. So we have

$$\Delta N-\tfrac{1}{2}(D^pN^a)\Pi_{ap}=0.$$

Substituting from (34) for D^pN^a, we get

$$\Delta N+\tfrac{1}{2}N\,\Pi^{ap}\Pi_{ap}=0$$

because $\epsilon^{apc}\Pi_{ap}=0$, Π_{ap} being symmetric. This Π is the Fischer–Marsden (Ref. 9) $\Pi-\frac{1}{2}(\operatorname{tr}\Pi)g$. So using the elliptic operator theory, $N=0$ unless $\Pi=0$ in which case N is constant. In the latter case (i.e., if $\Pi=0$), $F_{abc}=R_{abc}$ real. But then, Eq. (31″), after multiplying by $\tilde{\sigma}^d$ and taking Tr, gives

$$-\sqrt{2}iN\epsilon^{dbc}R_{abc}+N^bq^{cd}R_{abc}-\mathscr{D}_a\underline{N}^d=0.$$

Equating the imaginary part to zero and using

$$R_{abc}=(1/2\sqrt{2})R_{abpq}\epsilon^{pq}{}_c,$$

we get $NR_a{}^d=0$ or $R_{ad}=0$ if $N\neq0$, i.e., q_{ab} is flat.

But $\Pi=0$ and q is flat is ruled out by condition (2) of Theorem 1. Hence $\Pi\neq0$ and so $N=0$.

Substituting $N=0$ in (31′) and (31″), we get

$$\mathscr{D}_a(\mathrm{N}^{[a}\tilde{\sigma}^{b]})-\tfrac{1}{4}[\underline{N},\tilde{\sigma}^b]=0$$

and

$$N^bF_{ab}-\tfrac{1}{2}\mathscr{D}_a\underline{N}=0$$

or

$$\mathscr{L}_N\tilde{\sigma}^b-\tfrac{1}{4}[\underline{N},\tilde{\sigma}^b]=0,\quad-\mathscr{L}_NA_a-\tfrac{1}{2}\mathscr{D}_a\underline{N}=0.\tag{35}$$

But then condition (3) of the theorem give $N^a=0$ and the image of $\underline{N}$ lies in the center of $\mathscr{G}$. But then $\underline{N}$ is a function in the usual sense because it is invariant under gauge transformations (internal rotations) and thus globally defined. Thus if $N=0$, $N^a=0$ and the image of $\underline{N}$ lies in the center of $\mathscr{G}$, then (31″) gives $\mathscr{D}_a\underline{N}=0$ which now becomes $D_a\underline{N}=0$ or $\underline{N}=\text{const}$ in the center of $\mathscr{G}$, i.e., $\underline{N}=0$ in $(\Lambda^0\otimes\mathscr{G})^{\sim}$ where $(\Lambda^0\otimes\mathscr{G})^{\sim}$ is the quotient of $\Lambda^0\otimes\mathscr{G}$ by the constant functions with values in the center of $\mathscr{G}$. Here $(\Lambda^0\otimes\mathscr{G})^{\sim}$ is the L^2 dual of the space $\{\mathscr{D}_a\tilde{\sigma}\}$. Thus we have shown that $N=0$, $N^a=0$, and $\underline{N}=0$ in $(\Lambda^0\otimes\mathscr{G})^{\sim}$. Hence $D\Phi(\tilde{\sigma},A)^*$ is injective, and Theorem 1 is proved.

B. Group action and infinitesimal gauge covariance

Let P denote the principal bundle over ${}^4S$ with group $G=SU(2)$ [or $SL(2,C)$] acting on the right. Let $Q=P$ restricted to Σ. Given a local section i: $U\subset{}^4S\hookrightarrow P$, $\tau\circ i=\text{identity}$, where τ: $P\to{}^4S$ is the projection map. If $\tilde{\omega}$ denotes the connection one-form on P, then $A_\mu=i^*\tilde{\omega}$ and $F_{\mu\nu}=i^*\Omega$, where Ω is the curvature of $\tilde{\omega}$. (These are quantities needed in Ref. 3 to give Lagrangian formulation of Ashtekar's theory.) Here, $A_a=$ restriction of ${}^4A$ to $\Sigma=i^*$ ($\tilde{\omega}$ restricted to Q) and $F_{ab}=$ restriction of $F_{\mu\nu}$ to Σ.

Let $\mathscr{B}^3$ denote a semidirect product of the diffeomorphism group D^3 of Σ and the gauge transformation group G. The group G sits naturally in $\mathscr{B}^3$ and has the momentum map C_G. On the other hand, D^3 has no natural copy in $\mathscr{B}^3$. There is an action of D^3 on T^*C (C: extended configuration space) that is easily described in terms of its infinitesimal generators. An element of the Lie algebra of D^3 is a vector field N^a on Σ. For each point $(\tilde{\sigma},A)\in T^*C$, lift N^a horizontally to Q using the connection A. Let $\widetilde{N}^a$ denote this lifted vector field and $\bar{\sigma}$ be the tensorial object on Q corresponding to $\tilde{\sigma}$. Then the infinitesimal generator at $(\tilde{\sigma},A)$ of the $\mathscr{B}^3$ action on T^*C is given by $(\tau_Q^*(L_{\widetilde{N}}\bar{\sigma}),\tau_Q^*(L_{\widetilde{N}}(\tilde{\omega}|_Q)))$, where τ_Q is a local section of Q. This is precisely $(\mathscr{L}_N\tilde{\sigma},\mathscr{L}_NA)$, where

$$\begin{aligned}\mathscr{L}_N\tilde{\sigma}^b&=(D_aN^a)\tilde{\sigma}^b-\tilde{\sigma}^aD_aN^b+N^a\mathscr{D}_a\tilde{\sigma}^b\\&=(D_aN^a)\tilde{\sigma}^b-\tilde{\sigma}^aD_aN^b\\&\quad+N^aD_a\tilde{\sigma}^b+N^a[A_a,\tilde{\sigma}^b]\end{aligned}\tag{36}$$

and

$$\mathscr{L}_NA_a=N^bF_{ba}=N^b(\partial_aA_b-\partial_bA_a+[A_b,A_a]).\tag{36'}$$

These are the gauge-covariant Lie derivatives of Arms.[8] Also $\Theta=(C_a,C_G)$ is the momentum map for the action of $\mathscr{B}^3$ on T^*C.

Using (36) and (36′), verification of infinitesimal gauge covariance of $\Phi(\tilde{\sigma},A)=(C_G,C_a,C)$ under the action of $\mathscr{B}^3$ is quite straightforward:

$$\begin{aligned}\mathscr{L}_NC_G(\tilde{\sigma},A)&=DC_G(\tilde{\sigma},A)\cdot(\mathscr{L}_N\tilde{\sigma},\mathscr{L}_NA),\\\mathscr{L}_NC_a(\tilde{\sigma},A)&=DC_a(\tilde{\sigma},A)\cdot(\mathscr{L}_N\tilde{\sigma},\mathscr{L}_NA),\\\mathscr{L}_NC(\tilde{\sigma},A)&=DC(\tilde{\sigma},A)\cdot(\mathscr{L}_N\tilde{\sigma},\mathscr{L}_NA),\end{aligned}\tag{37}$$

or

$$\mathscr{L}_N\Phi(\tilde{\sigma},A)=D\Phi(\tilde{\sigma},A)\cdot(\mathscr{L}_N\tilde{\sigma},\mathscr{L}_NA).$$

Verification of the first equation of (37) needs the Jacobi identity of the Lie bracket. Calculations are straightforward and we omit them.

This infinitesimal covariance shows the fact that Einstein equations are covarient under diffeomorphisms of space-time. This is also an infinitesimal version of the fact that constraint equations themselves are covariant under bundle automorphisms.

The above result can now be used, as in FMM, Sec. 1, to show the gauge invariance of $D^2\Phi(\tilde{\sigma},A)$ and that of Taub's conserved quantity,

$$T=\int_\Sigma (N,N^a,\underline{N}_A{}^B)\cdot D^2\Phi(\tilde{\sigma},A)\cdot((\eta,B),(\eta,B)), \quad (38)$$

where $(N,N^a,\underline{N}_A{}^B)$ can be identified with the vector fields on the space-time bundle P by using Lemma 1 and the procedure in Sec. IV C of Ref. 8.

Proofs of the above statements are quite similar to those given in FMM, Sec. 1 and we omit them.

As remarked above, triplet $(N,N^a,\underline{N})\in$domain of $D\Phi^*$ can be identified as a vector field on the bundle. Also an analog of Moncrief's result holds: the kernel of $D\Phi(\tilde{\sigma},A)^*$ is isomorphic to the space of Killing fields (symmetries of the fields) that a space-time admits. This result can be proved by using Lemma 1 and the procedure used in Sec. IV C of Arms,[8] with obvious modifications and omissions in the notations.

Further, a nontrivial $(N,N^a,\underline{N})\in\ker D\Phi(\tilde{\sigma},A)^*$ gives rise to a second-order condition on linear perturbations (η,B),

$$\int_\Sigma (N,N^a,\underline{N})\cdot D^2\Phi(\tilde{\sigma},A)\cdot((\eta,B),(\eta,B))=0. \quad (39)$$

Again, proceeding as in Ref. 8, Secs. IV E and IV F, finally get the following analog of Theorem 2B of Ref. 8.

Theorem 2: The Einstein system is linearization stable at the solution ${}^4g$ if and only if there are no symmetries on the bundle P of the connection one-form $\tilde{\omega}$ except the action of the center of G on P. That is, if N^a is a vector field on P, such that $L_N\tilde{\omega}$ is zero, then N^a is a generator of the action of the center of G on P, i.e., N^a is vertical and $\tilde{\omega}(N^a)=$ const in the center of $\mathscr{G}$.

Recall that ${}^4g$ can also be recovered from ${}^4A$ and γ_a since through Lagrangian formulation (Ref. 3) σ can be known. Then σ and ${}^4A$ determine q_{ab} and the extrinsic curvature Π_{ab} on a hypersurface that determine ${}^4g$ through the usual Cauchy existence theorem up to space-time diffeomorphisms.

C. Moncrief's decomposition and the slice

Since $D\Phi(\tilde{\sigma},A)^*$ is elliptic, so is $-\mathbb{J}\circ D\Phi(\tilde{\sigma},A)^*$. Hence as in FMM, Sec. 2, we get

$$\begin{aligned}&T_{(\tilde{\sigma},A)}T^*C\\&\quad=\mathrm{Range}(-\mathbb{J}\circ D\Phi(\tilde{\sigma},A)^*)\oplus\mathrm{Range}(D\Phi(\tilde{\sigma},A)^*)\\&\qquad\oplus[\ker(D\Phi(\tilde{\sigma},A)\circ\mathbb{J})\cap\ker D\Phi(\tilde{\sigma},A)].\end{aligned}$$

The first summand represents the infinitesimal gauge transformations, the second summand is the orthogonal complement to the linearized constraints, and the last summand is the space of linearized "true" degrees of freedom, generalization of transverse-traceless quantities. The last summand preserves the constraints modulo gauge freedom.

The slice at $(\tilde{\sigma}_0,A_0)$ for the action of $\mathscr{B}^3$ is given by (as in AMM, Sec. 4) $S_0=\{(\tilde{\sigma}_0,A_0)\}+\mathscr{U}$, where $\mathscr{U}$ is a suitably small ball in $\ker D\Phi\circ\mathbf{J}$. For details, see FMM and Isenberg and Marsden.[12]

IV. SUFFICIENCY OF SECOND-ORDER CONDITIONS: CONICAL STRUCTURE OF SINGULARITIES

Here by singularities we mean those points in the set of solutions of Einstein's equations that are not linearization stable, i.e., which by Theorem 2, admit symmetries. Since results of linearized theory for Einstein's field equations are equivalent to those for constraint equations, we shall deal with constraint equations.

By arguments in the proof of Theorem 1, it follows that on Σ, a symmetry $(N,N^a,\underline{N})\in\ker D\Phi(\tilde{\sigma},A)^*$ satisfies either (a) $N=0$ (i.e., all symmetries are tangent to Σ) or (b) N is constant and the initial data are trivial, i.e., $\tilde{\sigma}$ corresponds to a flat metric and $\Pi=0$, $A=0$. In case (a), which is the "spacelike" case, there is a basis of $\ker D\Phi^*$ of the form

$$\{(0,X_i,V_i):\ i=1,2,\ldots,k,\ \mathscr{L}_{X_i}\tilde{\sigma}=\tfrac{1}{4}[V_i,\tilde{\sigma}]\ \text{and}\\ -\mathscr{L}_{X_i}A_a=\tfrac{1}{2}\mathscr{D}_aV_i\}.$$

In case (b), the timelike case, there is a basis with (1,0,0) as one element and the rest of the basis is as in case (a). Use of a Kuranishi map to get the conical structure of singularities in case (a) is exactly the same as AMM, Secs. 1 and 2 with obvious changes in notations. We briefly sketch this.

Let $\mathbb{P}$ denote the L^2-orthogonal projection to the range of $D\Phi(\tilde{\sigma}_0,A_0)$, $\mathbb{P}_\Theta$ denote the L^2-orthogonal projection onto the span of $(0,X_i,V_i)$, and $\mathbb{P}_H$ denote the projection onto (1,0,0). Thus the ith component of $\mathbb{P}_\Theta\circ\Phi$ is given by $\int_\Sigma(X_i^aC_a+\mathrm{Tr}(V_iC_G))$ and $\mathbb{P}_H\circ\Phi=\int_\Sigma C(\tilde{\sigma},A)$. It follows that if

$$\begin{aligned}\mathscr{C}_{\mathbf{P}}&=\{(\tilde{\sigma},A)/\mathbb{P}\circ\Phi(\tilde{\sigma},A)=0\},\\ \mathscr{C}_\Theta&=\{(\tilde{\sigma},A)/\mathbb{P}_\Theta\circ\Phi(\tilde{\sigma},A)=0\},\end{aligned}\quad (40)$$

and

$$\mathscr{C}_H=\{(\tilde{\sigma},A)/\mathbb{P}_H\circ\Phi(\tilde{\sigma},A)=0\},$$

then the space of solutions to constraint equations is

$$\mathscr{C}=\mathscr{C}_{\mathbf{P}}\cap\mathscr{C}_\Theta\ \text{ in case (a)}$$

and (41)

$$\mathscr{C}=\mathscr{C}_{\mathbf{P}}\cap\mathscr{C}_\Theta\cap\mathscr{C}_H\ \text{ in case (b)},$$

where $\mathbb{P}=\mathbb{P}_\Theta\oplus\mathbf{P}_H$. Then, using the slice described above, properties of the Kuranishi map, and gauge invariance of $D^2\Phi$, we get the following theorem giving the conical structure of symmetry solutions of constraint equations in case (a).

Theorem 3: The Kuranishi map F maps $\mathscr{C}_{\mathbf{P}}\cap\mathscr{C}_\Theta\cap S_0$ locally one to one and onto the cone,

$$\begin{aligned}C_\Theta=\{(\tilde{\sigma}_0,A_0)\}+\{(\eta,B)\in\ker D\Phi\cap\ker D\Phi\circ\mathbf{J}|\\ P_\Theta(\Phi-D\Phi)(\eta,B)=0\}.\end{aligned}$$

Thus removing gauges (slice conditions) as in FMM, this gives a conical structure of the solutions of constraint equations admitting spacelike Killing fields.

For the sake of convenience of the reader, we give a definition of a Kuranishi map and its properties used in the proof of Theorem 3. For the proof of these properties, see AMM, Sec. 1.

Kuranishi map: Let $(\tilde\sigma_0,A_0)\equiv x_0\in\Phi^{-1}(0)$ be fixed and let $\Delta = D\Phi(x_0)\circ D\Phi(x_0)^*$. Then by ellipticity of $D\Phi(x_0)^*$, Δ is an isomorphism of Range $D\Phi(x_0)$ to itself. Let $\mathbb{P}$ denote the orthogonal projection to range $D\Phi(x_0)$ and set $G=\Delta^{-1}\circ\mathbb{P}$, the Green's function for Δ. Write $y=x-x_0$ $[x\equiv(\tilde\sigma,A)]$ and let the remainder be given by $R(y)=\Phi(x)-D\Phi(x)\cdot y$. Define the Kuranishi map F by

$$F(x)=x+D\Phi(x_0)^*\circ G\circ R(y).$$

Then F satisfies the following properties.

(i) F is a diffeomorphism of a neighborhood of x_0 onto itself.

(ii) F maps the slice S_0 at x_0 to itself.

(iii) F is a local chart for $\mathscr{C}_{\mathbf{P}}$ and when restricted to $\mathscr{C}_{\mathbf{P}}\cap S_0$, F is a local symplectic diffeomorphism of $\mathscr{C}_{\mathbf{P}}\cap S_0$ to $\{x_0\}+(\ker D\Phi(x_0)\cap\ker D\Phi(x_0)\circ\mathbb{J})$.

(iv) The map of $\{x_0\}+\ker D\Phi(x_0)$ to $\mathscr{C}_{\mathbf{P}}$ given by $x_0+y\mapsto x_0+y+\Psi(y)$ is the inverse of the Kuranishi map when restricted to $\mathscr{C}_{\mathbf{P}}$. Here Ψ: $\ker D\Phi(x_0)\to$ Range $D\Phi(x_0)^*$ is a map defined on a neighborhood of zero such that $\Psi(0)=0$, $D\Psi(0)=0$, and such that $\mathscr{C}_{\mathbf{P}}$ is the graph of Ψ, i.e., locally

$$\mathscr{C}_{\mathbf{P}}=\{x=x_0+y+\Psi(y)\,|\,y\in\ker D\Phi(x_0)\}.$$

To treat case (b), we need some more technical details in addition to those needed in the proof of Theorem 3.

First, we need an analog of decomposition following Lemma 2.4 of AMM. Define

$$\sigma^{ab}=-\operatorname{Tr}\sigma^a\sigma_0^b$$

and (42)

$$A_{ab}=-\operatorname{Tr}A_a\sigma_0{}^b.$$

Recall that for case (b), σ_0 corresponds to a flat metric q_0 and $A_0=0=\Pi_0$. Since A_0 is the self-dual part of the Weyl tensor, it vanishes if the metric is flat. Then we prove the following.

Lemma 4: Elements of $\mathscr{C}_{\mathbf{P}}\cap S_0$ can be obtained as

$$\begin{aligned}\tilde\sigma^{ab}&=\tilde\sigma_0{}^{ab}+\gamma^{tt_{ab}}+(\tfrac{1}{3}\alpha q_0{}^{ab}+\sqrt{2}iD^bV^a)\sqrt{q_0},\\ A_{ab}&=B_{ab}{}^{tt}+\tfrac{1}{3}\beta q_{0_{ab}}+8\epsilon_{abp}D^pC\\ &\quad+4\sqrt{2}i(D_aY_b-(D_mY^m)q_{0_{ab}})+2iV^c\epsilon_{cab},\end{aligned}\qquad(43)$$

where $\gamma^{tt_{ab}}$ is a transverse traceless symmetric two-tensor density, $B^{tt}{}_{ab}$ is a transverse-traceless symmetric two-tensor, and $(C,Y,V)\in$ domain of $D\Phi^*$ is a function of γ^{tt}, α, B^{tt}, β.

Proof: We first compute $D\Phi(\tilde\sigma_0,0)^*\cdot(C,Y,V)$, where σ_0 corresponds to the flat metric q_0. In Eq. (31), put $A_0=0=F_{ab}$ and replace $(N,N^a,\underline{N})$ by (C,Y,V) to get

$$\begin{aligned}&D\Phi(\tilde\sigma_0,0)^*\cdot(C,Y,V)\\ &\quad=(\sqrt{2}i\mathscr{D}_bV,4\sqrt{2}\epsilon^{dbm}(D_dC)\tilde\sigma_{0m}\\ &\qquad+4\sqrt{2}i\mathscr{L}_Y\tilde\sigma_0{}^b-\sqrt{2}i[V,\tilde\sigma_0{}^b]).\end{aligned}\qquad(44)$$

Here we have used

$$[\tilde\sigma_0{}^a,\mathscr{D}_a\tilde\sigma_0{}^b]=i(\Pi_0{}^b{}_m-(\operatorname{tr}\Pi_0)\delta^b{}_m)\tilde\sigma_0{}^m\sqrt{q_0}=0,$$
$$\because\ \Pi_0=0.$$

Multiplying the second coordinate on the right-hand side of (44) by $-\sigma_0{}^a/\sqrt{q_0}$ and taking Tr, we get, after lowering the indices,

$$8\epsilon_{abp}D^pC+4\sqrt{2}i(D_aY_b-(D_mY^m)q_{0_{ab}})+2iV^C\epsilon_{cab}.\qquad(45)$$

Similarly, multiplying the first coordinate by $-\tilde\sigma_{0a}$, taking Tr, and raising the indices, we get

$$(\sqrt{2}iD^bV^a)\sqrt{q_0}.\qquad(45')$$

(These operations are in accord with the definition of $D\Phi^*$, its domain, and its range.)

We now use the fact that elements of ker $D\Phi(\tilde\sigma_0,0)\cap\ker D\Phi(\tilde\sigma_0,0)\circ\mathbb{J}$ consist of pairs

$$(\gamma^{tt_{ab}}+\tfrac{1}{3}\alpha q_0^{ab}\sqrt{q_0},B^{tt}_{ab}+\tfrac{1}{3}\beta q_{0ab}).$$

This can be proved by following Ashtekar.[13] Combining this fact, (45) and (45′), and the theory in AMM, Secs. 1 and 2, (or FMM, Secs. 5–7), we get the required decomposition (43) in the neighborhood of $(\tilde\sigma_0,0)$.

We also need an analog of Lemma 2.5 of AMM.

Lemma 5: If $\mathscr{L}_N\tilde\sigma_0=\tfrac{1}{4}[\underline{N},\tilde\sigma_0]$, then

$$\int_\Sigma(\mathscr{L}_N\tilde\sigma)^{ab}A_{ab}=\int_\Sigma(\mathscr{L}_N\gamma^{tt})^{ab}B^{tt}{}_{ab}.\qquad(46)$$

Proof: We first note that when there are no spinor indices, $\mathscr{L}_N$ is the ordinary Lie derivative and for an ordinary Lie derivative, the following identities hold:

$$\text{(I)}\ \int_\Sigma(L_Xh)^{ab}\omega_{ab}=-\int_\Sigma h^{ab}(L_Xw)_{ab},$$

w being a tensor density,

(II) if $D_bk^{ab}=0$ and $L_Xq_0=0$, then $D_b(L_Xk)^{ab}=0$.

We also note that by a simple calculation if $\mathscr{L}_N\tilde\sigma_0=\tfrac{1}{4}[\underline{N},\tilde\sigma_0]$, then $L_Nq_0^{ab}=0$.

Since we are working within $\mathscr{C}_{\mathbf{P}}\cap S_0$ $(S_0=S_{(\tilde\sigma_0,0)})$ we substitute decomposition (43) in the left-hand side of (46) and obtain (for brevity $\sqrt{q_0}=\mu_0$)

$$\begin{aligned}&\int_\Sigma(L_N\tilde\sigma)^{ab}A_{ab}\\ &\quad=\int_\Sigma L_N(\tilde\sigma^{ab}{}_0+\gamma^{tt_{ab}}+(\tfrac{1}{3}\alpha q^{ab}{}_0+\sqrt{2}iD^bV^a)\mu_0)\\ &\qquad\times(B^{tt}_{ab}+\tfrac{1}{3}\beta q_{0ab}+8\epsilon_{abp}D^pC\\ &\qquad+4\sqrt{2}i(D_aY_b-(D_mY^m)q_{0ab})+2iV^c\epsilon_{cab}).\end{aligned}$$

Since $\sigma_0{}^{ab}=q_0{}^{ab}$ and $L_Nq_0=0$, this becomes

$$I=\int_\Sigma L_N(\gamma^{tt_{ab}}+\sqrt{2}iD^bV^a\mu_0)\cdot\Box_{ab},$$

where

$$\square_{ab} = \text{second bracket in the above integral}$$
$$= B^{tt}_{ab} + \cdots + 2iV^c\epsilon_{cab}.$$

So,

$$I = -\int_\Sigma (\gamma^{tt_{ab}} + \sqrt{2}iD^bV^a\mu_0)\cdot L_N\square_{ab}. \qquad (47)$$

Now, by choice of the gauge $V^a = D^a V$ (which is valid since we are working within the slice), we see that $D^b\square_{ab} = 0$. Hence the term

$$\int_\Sigma D^bV^a\mu_0(L_N\square_{ab}) = -\int_\Sigma V^aD^b(L_N\square_{ab})\mu_0$$
$$= 0 \quad \text{by identity (II)}. \qquad (48)$$

See also remark 1 after Eq. (58) in Ref. 1 (first paper), where such a choice of gauge is justified for Hermiticity preserving evolution equations. So if one does not wish to use this gauge choice explicitly, one may work with Hermiticity preserving evolution equations, keeping all classical relativity intact.

Thus the remaining terms in (47) are

$$-\int_\Sigma \gamma^{tt_{ab}}L_N\square_{ab}$$
$$= -\int_\Sigma \gamma^{tt_{ab}}L_N(B^{tt}{}_{ab} + 8\epsilon_{abp}D^pC$$
$$+ 4\sqrt{2}i(D_aY_b - (D_mY^m)q_{0ab}) + 2iV^c\epsilon_{cab})$$
$$= \int_\Sigma (L_N\gamma^{tt})^{ab}(B^{tt}{}_{ab} + 8\epsilon_{abp}D^pC$$
$$+ 4\sqrt{2}i(D_aY_b - (D_mY^m)q_{0ab}) + 2iV^c\epsilon_{cab}). \qquad (49)$$

Since $L_N\gamma^{tt_{ab}}$ is symmetric in a,b, we get at once

$$(L_N\gamma^{tt})^{ab}\epsilon_{abp}D^pC = 0 \qquad (50)$$

and

$$(L_N\gamma^{tt})^{ab}\epsilon_{cab}V^c = 0. \qquad (51)$$

Next, consider

$$\int_\Sigma L_N\gamma^{tt_{ab}}D_aY_b = -\int_\Sigma Y_bD_a(L_N\gamma^{tt})^{ab} = 0, \qquad (52)$$

because $D_a\gamma^{tt_{ab}} = 0$ and so by identity (II),

$$D_a(L_N\gamma^{tt})^{ab} = 0.$$

Lastly, consider $\int_\Sigma (L_N\gamma^{tt})^{ab}(D_mY^m)q_{0ab}$. Here

$$q_{0ab}(L_N\gamma^{tt})^{ab}$$
$$= q_{0ab}(N^kD_k\gamma^{tt_{ab}} - \gamma^{tt_{ak}}D_kN^b - \gamma^{tt_{bk}}D_kN^a$$
$$+ (D_kN^k)\gamma^{tt_{ab}})$$
$$= N^kD_k(q_{0ab}\gamma^{tt_{ab}}) - (\gamma^{tt}{}_{bk}D^kN^b + \gamma^{tt}{}_{kb}D^bN^k)$$
$$+ (D_kN^k)q_{0ab}\gamma^{tt_{ab}}$$
$$= -\gamma^{tt}{}_{bk}(D^kN^b + D^bN^k) = -\gamma^{tt}{}_{bk}L_Nq_0^{bk} = 0$$
$$(53)$$

since $q_{0ab}\gamma^{ab} = 0$ and $L_Nq_0 = 0$. Thus combining (50)–(53) with (49), the only term left is the right-hand side of (46).

Note: Relations like

$$\int_\Sigma k^{ab}D_bV_a = -\int_\Sigma V_aD_bk^{ab}$$

are valid after integration by parts because Σ is compact and all quantities involved are elements of suitable Sobolev spaces and hence distributions with compact support, so that boundary terms vanish.

We are now in a position to treat the timelike case, the scalar (Hamiltonian) constraint. We have to study the intersection $\mathscr{C}\cap S_0 = \mathscr{C}_H\cap\mathscr{C}_\Theta\cap\mathscr{C}_P\cap S_0$. From Lemma 4 above and Lemma 1.5 of AMM, C, Y, V are smooth mappings of $\gamma^{tt}, B^{tt}, \alpha, \beta$ that together with their first derivatives vanish at (0, 0, 0, 0) and have the following property: For any α and β and any (γ^{tt}, B^{tt}) satisfying

$$\int_\Sigma (L_{N_i}\gamma^{tt})^{ab}B^{tt}_{ab} = 0, \qquad (54)$$

where $N_1,N_2,\dots,N_k$ are Killing fields of q_0 (or satisfy $\mathscr{L}_{N_i}\tilde{\sigma}_0 = \frac{1}{4}[\underline{N}_i\tilde{\sigma}_0]$), the data $(\tilde{\sigma}^{ab},A_{ab})$ given by (43) lie in $\mathscr{C}_P\cap\mathscr{C}_\Theta\cap\bar{S}_0$. Thus the mappings (C,Y,V) parametrize a full neighborhood of $(\tilde{\sigma}_0,0)$ in $\mathscr{C}_P\cap\mathscr{C}_\Theta\cap S_0$ in terms of solutions (γ^{tt},B^{tt}) of (54) and (α,β). The cone given by (54) restricts (γ^{tt},B^{tt}) but leaves (α,β) unrestricted. Now consider an affine submanifold of T^*C:

$$M = \{(\tilde{\sigma}_0{}^a + \gamma^a,0)\in T^*C \,|\, \mathscr{D}_b\gamma^a = 0\}.$$

We claim that each point of M is a critical point of the function f: $(\tilde{\sigma},A)\mapsto\int_\Sigma [C(\tilde{\sigma},A)/\mu_0]$ and $\int_\Sigma [C(\tilde{\sigma},A)/\mu_0]$ vanishes on M. For this, we want to prove that $Df(\tilde{\sigma},A)\cdot(\eta,B) = 0$, $\forall(\tilde{\sigma},A)\in M$. Now,

$$Df(\tilde{\sigma},A)\cdot(\eta,B) = D\left(\int_\Sigma \frac{C(\tilde{\sigma},A)}{\mu_0}\right)\cdot(\eta,B)$$
$$= \int_\Sigma DC(\tilde{\sigma},A)\cdot(\eta,B)\frac{1}{\mu_0}$$
$$= \int_\Sigma \left\langle\left(\frac{1}{\mu_0},0,0\right),D\Phi(\tilde{\sigma},A)\cdot(\eta,B)\right\rangle$$
$$= \int_\Sigma \left\langle D\Phi(\tilde{\sigma},A)^*\cdot\left(\frac{1}{\mu_0},0,0\right),(\eta,B)\right\rangle.$$
$$(55)$$

Now $\tilde{\sigma}_0$ corresponds to flat q_0 and since $\mathscr{D}_a\gamma^b = 0$, γ also corresponds to flat q because $A = 0$. So $\tilde{\sigma}_0 + \gamma$ corresponds to flat q, hence corresponding A_a and F_{ab} are both zero. Hence by (44),

$$D\Phi(\tilde{\sigma},0)^*\cdot(1/\mu_0,0,0)$$
$$= (0,4\sqrt{2}\epsilon^{dbm}D_b(1/\mu_0)\tilde{\sigma}_m) = (0,0).$$

So, $Df(\tilde{\sigma},0)\cdot(\eta,B) = 0$, for every $(\tilde{\sigma},0)\in M$, i.e., $Df(\tilde{\sigma},0) = 0$, $\forall(\tilde{\sigma},0)\in M$.

Thus each point of M is a critical point of $\int_\Sigma C(\tilde{\sigma},A)/\mu_0$. Hence $D^2f(\tilde{\sigma}_0,0)$ is well defined. The degeneracy space of $D^2f(\tilde{\sigma}_0,0)$ is exactly $T_{(\tilde{\sigma}_0,0)}M$.

Arguments above also show that $\int_\Sigma [C(\tilde{\sigma},A)/\mu_0]$ vanishes whenever $(\tilde{\sigma},A)\in M$.

Now

$$C(\tilde{\sigma},A) = -2\,\mathrm{Tr}\,\tilde{\sigma}^a\tilde{\sigma}^b F_{ab}$$
$$= -2\epsilon^c{}_{mn}\tilde{\sigma}^{am}\tilde{\sigma}^{bn}(-\partial_b A_{ac}$$
$$+\partial_a A_{bc} + \sqrt{2}\epsilon^{pq}{}_c A_{ap}A_{bq}). \qquad (56)$$

Next step is to substitute decompositions (43) for $\tilde{\sigma}^{ab}$ and A_{ab} in (56) and consider $\int_\Sigma [C(\tilde{\sigma},A)/\mu_0]$. This gives

$$\int_\Sigma \frac{C(\tilde{\sigma},A)}{\mu_0} = \int_\Sigma \mu_0 B^{tt_{ab}} B^{tt}{}_{ab} + \int_\Sigma \frac{\gamma^{tt_{ab}}\gamma^{tt}{}_{ab}}{\mu_0}$$
$$-6\beta^2\,\mathrm{vol.}(\Sigma) + G(\gamma^{tt},\alpha,B^{tt},\beta). \quad (57)$$

Here $B^{ab}B_{ab}$ contains terms $w^{ab}w_{ab} + \nabla\gamma\cdot\nabla\gamma$ and so (57) tallies with expression in Lemma 3.1 of AMM. Combining all above considerations, we get the following lemma (analog of Lemma 3.1 of AMM).

Lemma 6: In a neighborhood of $(\tilde{\sigma}_0,0)$ we have Eq. (57) where first and second derivatives of G vanish at (0,0,0,0). Also, each point of M is a critical point of $\int_\Sigma [C(\tilde{\sigma},A)/\mu_0]$ and $\int_\Sigma [C(\tilde{\sigma},A)/\mu_0]$ vanishes on M. Also, G vanishes on M and so do its first and second derivatives. Here M is a nondegenerate critical manifold for $\int_\Sigma [C(\tilde{\sigma},A)/\mu_0]$ in the sense of FMM, Sec. 6. Here $(\tilde{\sigma},A)$ is to be regarded as a function of variables γ^{tt}, α, B^{tt}, β. Without imposing (54), we have to substitute (43) in $\int_\Sigma (1/\mu_0)C(\tilde{\sigma},A)$. Thus $\int_\Sigma (1/\mu_0)C(\tilde{\sigma},A)$ is a smooth function of γ^{tt}, α, B^{tt}, β and we have to consider its Taylor expansion in these variables around (0,0,0,0).

The remaining procedure is then as in AMM, Sec. 3 and we get a cone $C_{\circledcirc}$ defined by (54) in variables $(\bar{\gamma}^{tt}, \bar{\alpha}, \bar{B}^{tt}, \bar{\beta})$ obtained from $(\gamma^{tt}, \alpha, B^{tt}, \beta)$ by a change of coordinates through the parametrized Morse lemma to eliminate the higher-order terms. Cone C_H is defined by

$$\bar{\beta}_\pm = \pm\frac{1}{6}\left[\int_\Sigma (B^{tt_{ab}}B^{tt}{}_{ab})\mu_0 + \int_\Sigma \frac{1}{\mu_0}\gamma^{tt_{ab}}\gamma^{tt}{}_{ab}\right] \qquad (58)$$

with two branches (±). This corresponds to the scalar constraint.

Thus the cone $C_{\Theta}\cap C_H\cap S_0$ consists of those $(\gamma^{tt},\alpha,B^{tt},\beta)$ such that (54) holds and

$$\int_\Sigma (B^{tt_{ab}}B^{tt}_{ab})\mu_0 - 6\beta^2\,\mathrm{vol.}\,\Sigma + \int_\Sigma \frac{1}{\mu_0}\gamma^{tt_{ab}}\gamma^{tt}{}_{ab} = 0. \qquad (59)$$

Thus we get the final theorem.

Theorem 5: The association $(\gamma^{tt},\alpha,B^{tt},\beta)\mapsto(\tilde{\sigma},A)$, where $(\tilde{\sigma},A)$ are given by (43) with $\beta = \beta_\pm\ (\gamma^{tt},\alpha,B^{tt},\beta)$ defined by (58) with ± depending on the sign of β, is a one to one correspondence between the cone $C_{\circledcirc}\cap C_H\cap S_0$ defined by (54) and (59) and the nonlinear constraint set $\mathscr{C}\cap S_0$ in a neighborhood of $(\tilde{\sigma}_0,0)$. This correspondence maps straight lines in the cone through $(\tilde{\sigma}_0,0)$ (i.e., a solution of the linearized equations satisfying the second-order conditions) to a smooth curve in $\mathscr{C}\cap S_0$ with the same tangent at $(\tilde{\sigma}_0,0)$.

Hence second-order conditions on linearized perturbations are sufficient for the existence of an exact perturbation curve.

Gauge conditions can be removed by eliminating S_0 as explained in FMM. See also Isenberg and Marsden.[12]

V. DISCUSSION AND FURTHER WORK

Results obtained above can also be derived for systems coupled to Einstein equations, such as Einstein–Maxwell, Einstein–Yang–Mills, and Einstein–Klein–Gordon with the Einstein part treated through new variables (see Arms,[8,14] AMM, and Saraykar[15] for results in the ADM formalism). Especially, preliminary calculations show that treatment of the Einstein–Yang–Mills system is simplified due to similarity in the expressions for momentum constraints in new variables and those of Yang–Mills. Also, the geometrical setting remains the same. Calculations are not very different than those presented here and the Fischer–Marsden–Moncrief program runs quite along the same lines except for additional terms in the Yang–Mills variables.

In future, we propose to give a new variables approach for asymptotically flat space-times capturing asymptotic behavior of the variables in suitable function spaces, say, Choquet–Bruhat–Christodoulou weighted Sobolev spaces[16] and discuss linearization stability for these space-times. Following an unpublished result of Ashtekar,[17] we intend to prove that such space-times are always linearization stable, whether they admit symmetries or not.

As in AMM, in the case of a compact Cauchy hypersurface, it follows that the space of solutions modulo bundle automorphisms over diffeomorphisms is a stratified symplectic manifold, i.e., a stratified manifold, each stratum of which is symplectic. In the future, we wish to extend the York analysis to new variables.[18] Using the slice theorem and this York analysis, it can then be proved (cf. Ref. 12) that the generic points consisting of space-times with no symmetries are an open and dense set. Thus the generic symplectic stratum in the reduced space is also open and dense.

ACKNOWLEDGMENTS

I express my deep gratitute to Professor A. Ashtekar for introducing me to his new variables approach to Hamiltonian formulation of general relativity, for his encouragement to work out the FMM program via new variables, and for many suggestions in the beginning of this work. I also thank Dr. Naresh Dadhich for enabling me to attend Professor Ashtekar's lectures on "new variables" at Poona University.

[1] A. Ashtekar, "A new Hamiltonian formulation of general relativity," Syracuse University preprint, 1986; see also, A. Ashtekar, Phys. Rev. Lett. **57**, 2244 (1986) and A. Ashtekar, in *Proceedings of the Florence Conference on "Constrained Systems"* (World Scientific, Singapore, 1986).

[2] P. Renteln, ITP-UCSB preprint, 1986.

[3] J. Samuel, "A Lagrangian basis for Ashtekar's reformulation of canonical gravity," preprint, Raman Research Institute, Bangalore, 1987.

[4] A. Fischer, J. Marsden, and V. Moncrief, Ann. Inst. H. Poincaré **33**, 147 (1980).

[5] J. Arms, J. Marsden, and V. Moncrief, Ann. Phys. (NY) **144**, 81 (1982).

[6] T. Jacobson and L. Smolin, ITP-USCB preprint, 1986.

[7] P. Renteln and L. Smolin, ITP-UCSB preprint, 1986.

[8] J. Arms, J. Math. Phys. **20**, 443 (1979).

[9] A. Fischer and J. Marsden, in *Proceedings of the International School of Physics, "E. Fermi,"* course LXVII, *Isolated Gravitating Systems in General Relativity*, edited by J. Ehlers (Academic, New York, 1979), p. 322.

[10] R. Arnowitt, S. Deser, and C. Misner, in *Gravitation, An Introduction to Current Research*, edited by L. Witten (Wiley, New York, 1962).

[11] A. Sen, J. Math. Phys. **22**, 1781 (1981).

[12] J. Isenberg and J. Marsden, Phys. Rep. **89**, 179 (1982).
[13] A. Ashtekar, lectures given at Poona University, India, December 1986.
[14] J. Arms, J. Math. Phys. **18**, 830 (1977).
[15] R. Saraykar, Pramana **20**, 293 (1983).
[16] See, for example, Y. Choquet-Bruhat and D. Christodoulou, Acta Math. **146**, 129 (1981).
[17] A. Ashtekar (personal communication).
[18] For York analysis in the ADM formulation, see for example, M. Cantor, Commun. Math. Phys. **57**, 83 (1977); J. Isenberg and J. Marsden, J. Geom. Phys. **1**, 85 (1984).

Gravitational repulsion in sources of the Reissner–Nordström field

J. Ponce de Leon[a)]

Universidad Simon Bolivar, Division de Fisica y Matematicas, Departamento de Fisica, Apdo. 80659, Caracas 1081-A, Venezuela and Departamento de Fisica, Facultad de Ciencias, Universidad Central de Venezuela Caracas 1051, Venezuela

(Received 24 December 1986; accepted for publication 13 August 1987)

A number of aspects of the phenomenon of gravitational repulsion in static sources of the Reissner–Nordström field are investigated. It is found that in the case of perfect fluid spheres there exists a close relation between this phenomenon and the Weyl curvature tensor. In fact, it is proved that such a source gives rise to gravitational repulsion only if the pure gravitational field energy inside the sphere is negative. It is also proved that although the gravitational repulsion always takes place in the interior of a perfect fluid charged sphere when its radius r_0 is less than the "classical electron radius" r_e, this is not necessarily so either in the case of anisotropic charged spheres or if the net charge of the body is concentrated at its boundary only. It is shown that the phenomenon can also occur inside charged sources with $r_0 > r_e$. New sources of repulsive gravitation are constructed. They differ from others in the literature, since they neither satisfy the equation of state of "false vacuum" nor are their total gravitational masses entirely of electromagnetic origin. It is found that the charge contributes negatively to the effective gravitational mass M_G, in the sense that an increase in the charge causes a decrease in M_G. The gravitational repulsion in the new models constructed here is explained as due to this negative contribution rather than due to the strain of vacuum because of vacuum polarization.

I. INTRODUCTION

This paper deals with the phenomenon of gravitational repulsion in static sources of the Reissner–Nordström field. This phenomenon has recently been discussed in the literature and appears in regions where the effective gravitational mass M_G, which influences the motion of test particles, becomes negative.[1–3]

In order to describe the aspects of the phenomenon with which we deal with in this work let us briefly summarize some previous results. The effective gravitational mass (including the gravitational field energy) inside a volume V is given by the Tolman–Whittaker formula, viz.,

$$M_G = \int_V (T^0_0 - T^1_1 - T^2_2 - T^3_3)\sqrt{-g}\, dV, \tag{1.1}$$

where the T^μ_ν are the components of the energy momentum tensor of the matter. In the region outside of a spherically symmetric charged source Eq. (1.1) has been evaluated by Cohen and Gautreau[4] as

$$M_G(r) = M(1 - r_e/r) \tag{1.2}$$

with

$$r_e = q^2/M, \tag{1.3}$$

where q and M are the charge and the mass of the source, respectively. Notice that r_e in Eq. (1.3) is equal to the classical electron radius when q and M are the charge and the mass of the electron. Therefore r_e is called the "electron radius" here. Equations (1.2) and (1.3) show that the effective gravitational mass outside the source decreases from M to zero as r decreases from infinity to $r = r_e$ and becomes negative for $r < r_e$. Sources with $M^2 > q^2$ have the horizon at $r_+ = M + (M^2 - q^2)^{1/2}$ and a null surface at $r_- = M - (M^2 - q^2)^{1/2}$, so that $r_+ > M > r_e$. If $M^2 = q^2$ then $r_+ = r_- = r_e = M$. Consequently as long as one is outside a source with $M^2 \geqslant q^2$ and outside the even horizon, M_G is positive. However, in the case of sources with $q^2 > M^2$ there is no horizon and the radius of a static source (say r_0) can be less than the classical electron radius, resulting in a repulsion of uncharged test particles that approach the neighborhood of such sources. Examples of this kind are the models of Tiwari, Rao, and Kanakamedala (TRK).[5] In these models the radius of the source is 4/5 times the classical electron radius r_e and M_G is negative, not only in the vicinity of the source, but also at all interior points [the latter is calculated from Eq. (1.1), see Ref. 1]. Gautreau[2] has constructed a spherically symmetric electron model, so the radius of the source is equal to r_e. Consequently, outside the source M_G is positive and there is no gravitational repulsion around the "particle." However, a simple calculation shows that the effective gravitational mass M_G, as given by Eq. (1.1), is negative inside Gautreau's electron which implies that the interior of the body is gravitationally repulsive.

Other sources of repulsive gravitation have been investigated by Grøn.[1] In particular he studied the solution of Cohen and Cohen[6] and interpreted it as representing a charged spherical shell whose interior is described by the static form of the de Sitter metric. Moreover he showed that the blueshift of light propagating from the center in this model is produced by the negative effective gravitational mass M_G inside the shell.

Important common features of all the above mentioned sources of repulsive gravitation are the following: first, that they obey the equation of state of "false vacuum," viz., $\rho = -p, \rho > 0$; and second, that they are models wherein all

[a)] Present address: Apartado 2816, Caracas 1010-A, Venezuela.

0022-2488/88/010197-10$02.50

physical quantities, such as the energy density, pressure, and mass, are dependent on the charge alone and vanish when the charge vanishes; in other words, they are models of bodies whose mass is completely of electromagnetic origin.

As we see the phenomenon of gravitational repulsion in sources of the Reissner–Nordström field has been so far discussed on the basis of a very special (and restricted) class of solutions of the Einstein–Maxwell equations. In this paper we want to discuss a number of aspects of the phenomenon that arise from the consideration of the above mentioned results. We specifically refer to the following questions.

(1) What are the necessary conditions for the radius of a charged source r_0 to be less than r_e?

(2) Is the interior of a charged source with $r_0 < r_e$ in all cases gravitationally repulsive?

(3) Can there exist gravitational repulsion in sources whose radius $r_0 > r_e$?

(4) Can there exist gravitational repulsion in charged sources not satisfying the equation of state of false vacuum?

(5) Is the total mass (i.e., the mass measured by an observer at infinity) in all models that produce gravitational repulsion always of electromagnetic origin?

In the case where the net charge resides entirely at the boundary of the body we ask the following questions.

(6) How does the incorporation of charge, on the surface of the body, modify the behavior of the source regarding gravitational repulsion? If in the absence of charge there is no gravitational repulsion, then can the interior of the sphere become repulsive due to the surface charge?

Each of the above questions will be treated in some detail in Secs. II–IV. In Sec. II, without invoking any specific interior solution, we study some properties of charged spheres. We use the Weyl tensor, which describes the free gravitational field, to derive our answers to questions (1)–(3). Moreover we generalize the discussion by introducing anisotropic fluids. The motivation for this generalization is the following: In the case of matter without charge it is known that the properties of anisotropic sources may differ drastically from the properties of isotropic,[7–11] and it is therefore quite natural to expect that also in the charged case the introduction of anisotropy may lead to distributions with interesting physical properties. In particular we will see that although under some circumstances the radius of a perfect fluid source always is larger than r_e (and therefore there is no gravitational repulsion around the source), this is not necessarily so when the source has anisotropic "pressures." We will also see that the introduction of anisotropy allows us to construct models of charged bodies that have no perfect fluid analog. However, we do not discuss here the mechanism inducing a possible anisotropy. Rather we are interested in knowing to what extent the anisotropic sources may differ from the isotropic ones regarding the phenomenon of gravitational repulsion.

To answer questions (4) and (5) we need to find additional sources of repulsive gravitation distinct from those of TRK and of Gautreau. We undertake such a task in Sec. III wherein motivated by physical considerations we construct a four-parameter solution of the Einstein–Maxwell equations that contains (for specific values of the parameters) some previously known solutions in the literature. We analyze the properties of our models and show specific distributions that elucidate questions (4) and (5).

In Sec. IV we consider the neutral version of the solution constructed in Sec. III under the additional assumption of surface concentration of charge. The solution thus obtained is used to study the answer to question (6). The relation between our model to other models in the literature is also discussed. Our results are summarized in Sec. V.

II. SOURCES OF THE REISSNER-NORDSTRÖM FIELD AND GRAVITATIONAL REPULSION

A. Field equations

Let us consider a static distribution of matter represented by a charged spherically symmetric fluid which may be anisotropic.

In Schwarzschild coordinates the line element assumes the form

$$ds^2 = e^{\nu(r)}\,dt^2 - e^{\lambda(r)}\,dr^2 - r^2(d\theta^2 + \sin^2\theta\, d\phi^2). \tag{2.1}$$

With this choice of coordinates the energy-momentum tensor is diagonal,[12] viz.,

$$T^{\mu}_{\nu} = \mathrm{diag}\Big(\rho + \frac{E^2}{8\pi}, -p_r + \frac{E^2}{8\pi}, -p_\perp - \frac{E^2}{8\pi}, -p_\perp - \frac{E^2}{8\pi}\Big), \tag{2.2}$$

$$T^{\mu}_{\nu} = 0, \quad \mu \neq \nu,$$

where ρ is the energy density of matter, E is the usual electric field intensity, and p_r and $p_\perp$ are, respectively, the radial and tangential pressure which may be unequal.

The Einstein–Maxwell equations may be written as

$$8\pi T^0_0 = 8\pi\rho + E^2 = -e^{-\lambda}\Big(\frac{1}{r^2} - \frac{\lambda'}{r}\Big) + \frac{1}{r^2}, \tag{2.3}$$

$$8\pi T^1_1 = -8\pi p_r + E^2 = -e^{-\lambda}\Big(\frac{1}{r^2} - \frac{\nu'}{r}\Big) + \frac{1}{r^2}, \tag{2.4}$$

$$8\pi T^2_2 = 8\pi T^3_3 = -8\pi p_\perp - E^2 = -\frac{e^{-\lambda}}{2}\Big(\nu'' + \frac{\nu'^2}{2} + \frac{\nu' - \lambda'}{r} - \frac{\nu'\lambda'}{2}\Big), \tag{2.5}$$

where the prime denotes differentiation with respect to r. The electric field is

$$E = \frac{4\pi}{r^2}\int_0^r r^2\rho_e\, dr \equiv \frac{q(r)}{r^2}, \tag{2.6}$$

where $q(r)$ is the charge inside a sphere of "radius" r and ρ_e is the charge density which is related to the proper charge density $\bar{\rho}_e$ by

$$\rho_e = \bar{\rho}_e e^{\lambda/2}. \tag{2.7}$$

The space-time exterior to the source is described by the Reissner–Nordström field which in curvature coordinates has the form

$$ds^2 = f^2\,dt^2 - f^{-2}\,dr^2 - r^2(d\theta^2 + \sin^2\theta\, d\phi^2) \tag{2.8}$$

with

$$f^2(r) = 1 - \frac{2M}{r}\left(1 - \frac{r_e}{2r}\right) = 1 - \frac{(M + M_G)}{r}.$$

B. Some general relations: Perfect fluid case

We now proceed to show some general relations for static sources of the Reissner–Nordström field arising from the fulfillment of the boundary conditions.

The Einstein–Maxwell equation (2.3) can be written as

$$e^{-\lambda(r)} = 1 - \frac{8\pi}{r}\int_0^r \left(\rho + \frac{E^2}{8\pi}\right) r^2\, dr, \tag{2.9}$$

the continuity of $e^{-\lambda}$ across the boundary $r = r_0$ yields

$$\left(1 - \frac{r_e}{2r_0}\right) = \frac{4\pi}{M}\int_0^{r_0} \left(\rho + \frac{E^2}{8\pi}\right) r^2\, dr. \tag{2.10}$$

From this equation we obtain the following.

(a) If $\rho > 0$, then

$$r_0 > r_e/2, \tag{2.11}$$

a result which has been previously obtained by Bonnor[13] and by Tod.[14]

(b) If the charge is concentrated in a thin shell with zero proper energy density, then

$$r_0 = r_e/2. \tag{2.12}$$

(c) **Theorem:** If $\rho > 0$ and T^0_0 does not increase outward, then in the absence of surface concentration of charge at $r = r_0$,

$$r_0 \geqslant 2r_e/3. \tag{2.13}$$

Proof: The condition $(T^0_0)' \leqslant 0$ implies that

$$[8\pi\rho(r) + E^2(r)] \geqslant [8\pi\rho(r_0) + E^2(r_0)].$$

Substituting this expression into Eq. (2.10) and using that $E(r_0) = q/r_0^2$ we obtain (2.13).

Remark: We stress the fact that the limiting value $r_0 = 2r_e/3$ is attainable by gaseous sources in the case of $T^0_0 = \text{const}$ only.[15]

Now, we want to find some information about the dependence of the size of the source upon other physical quantities. With this aim we consider the decomposition of the Reimann tensor into the conformal Weyl tensor $C_{\mu\nu\alpha\beta}$, Ricci tensor, and its spur,[16] viz.,

$$R_{\mu\nu\alpha\beta} = C_{\mu\nu\alpha\beta} + \tfrac{1}{2}R_{\mu\alpha}g_{\nu\beta} - \tfrac{1}{2}R_{\nu\alpha}g_{\mu\beta} + \tfrac{1}{2}R_{\nu\beta}g_{\mu\alpha} - \tfrac{1}{2}R_{\mu\beta}g_{\nu\alpha} - (R/6)[g_{\mu\alpha}g_{\nu\beta} - g_{\mu\beta}g_{\nu\alpha}]. \tag{2.14}$$

The Weyl tensor, which has all the symmetry properties of the Riemann tensor, is to be thought of as representing the free gravitational field.

Using the field equations (2.3)–(2.5) in (2.14) we obtain

$$m = (4\pi/3)r^3 T^0_0 + (4\pi/3)r^3(T^1_1 - T^2_2) + W, \tag{2.15}$$

where m and W are defined by

$$m = rR^3{}_{232}/2, \quad W = rC^3{}_{232}/2; \tag{2.16}$$

here m is the usual mass function,[17] viz., $m = (1 - e^{-\lambda})r/2$ and $W(r)$, which represents the contribution of the free gravitational field, can be interpreted as the purely gravitational field energy inside a sphere of radius r.[16,18,19]

As an immediate consequence of Eq. (2.15) one has the following.

(d) **Theorem:** In the case of charged perfect fluid sources with vanishing density ρ at the boundary, the necessary and sufficient condition for $r_0 \leqslant r_e$ is that $W(r_0) \leqslant 0$.

Proof: Using Eqs. (2.2) and (2.10) to evaluate (2.15) at the boundary $r = r_0$ of a perfect fluid sphere we obtain

$$M(1 - r_e/r_0) = W(r_0) + (4\pi/3)r_0^3\rho(r_0). \tag{2.17}$$

From this equation we see that if $\rho(r_0) = 0$, then $r_0 \leqslant r_e \Leftrightarrow W(r_0) \leqslant 0$, which concludes the proof.

(d1) *Corollary:* If the contribution of the free gravitational field to the mass of the body is positive, viz., $W(r_0) > 0$, then the radius of a perfect fluid source always is larger than r_e.

(d2) *Corollary:* If the radius of a perfect fluid source is less than r_e, then necessarily $W(r_0) < 0$.

(d3) *Corollary:* For charged perfect fluid sources in a conformally flat space-time ($W = 0$), $r_0 \geqslant r_e$ ($r_0 = r_e$ only for gaseous spheres).

(e) **Theorem:** If $W(r) \leqslant 0$ throughout a perfect fluid charged sphere with $r_0 \leqslant r_e$ then all the interior points of such a sphere are gravitationally repulsive.

Proof: The effective gravitational mass M_G inside a sphere of radius r is given by the Tolman–Whittaker formula as

$$M_G(r) = 4\pi\int_0^r (T^0_0 - T^1_1 - T^2_2 - T^3_3)r^2 e^{(\nu+\lambda)/2}\, dr. \tag{2.18}$$

Using the field equations (2.3)–(2.5), Eq. (2.18) may be written as

$$M_G(r) = \tfrac{1}{2}r^2 e^{(\nu-\lambda)/2}\nu'. \tag{2.19}$$

Subtracting Eq. (2.5) from (2.4) and substituting, in the result, ν' and ν'' as obtained from Eq. (2.19), we get

$$rM'_G - 3M_G = e^{(\nu+\lambda)/2}\{r^3[4\pi(p_\perp - p_r) + E^2] - 3W\}. \tag{2.20}$$

The integration of this expression yields

$$M_G(r) = M\left(\frac{r}{r_0}\right)^3\left(1 - \frac{r_e}{r_0}\right) + r^3\int_r^{r_0} \frac{e^{(\nu+\lambda)/2}}{r^4} \times\{3W - r^3[4\pi(p_\perp - p_r) + E^2]\}dr. \tag{2.21}$$

We see that in the case of perfect fluid ($p_r = p_\perp$) both terms in the right-hand side are nonpositive when $r_0 \leqslant r_e$ and $W(r) \leqslant 0$. Consequently in this case $M_G < 0$ and the body is gravitationally repulsive at every point.

(e1) *Corollary:* A gaseous perfect fluid charged sphere, in a conformally flat space-time, is gravitationally repulsive at all its interior points.

Proof: The proposition follows directly from (d3) and Theorem (e).

(e2) *Corollary:* The phenomenon of gravitational repulsion always takes place in the interior of charged perfect fluid spheres if $r_0 \leqslant r_e$.

Proof: What we have to show is that M_G always becomes negative in such spheres. This is evident when $r_0 < r_e$ since from Eq. (2.21) we see that M_G is negative near the boundary not only for isotropic spheres but also for anisotropic ones. However, the proposition is not evident when $r_0 = r_e$ [see point (iii) in Sec. II C]. In the latter case and for perfect fluid we can expand the Eq. (2.18) near r_0 and use the boundary conditions to obtain

$$M_G(r) \approx -4\pi r_0^2 [\rho(r_0) + q^2/4\pi r_0^4](r_0 - r), \quad (2.22)$$

from which it follows that $M_G < 0$ near the boundary when $r_0 = r_e$. Finally we notice from Eq. (2.21) that the phenomenon of gravitational repulsion could also occur inside perfect fluid sources even when $r_0 > r_e$ if, for example, the purely gravitational field energy W takes sufficiently large negative values.

C. Anisotropic matter

In the discussion of points (a)–(e) we have used, basically, the boundary conditions, viz., the continuity of the metric functions, the continuity of the electric field, and that the pressure p ($= p_r = p_\perp$) vanishes at the boundary. In the case of anisotropic fluid the boundary conditions differ from those of perfect fluid in that only the radial pressure p_r must vanish at the boundary r_0. In general the tangential pressure $p_\perp$ (as well as ρ) may be discontinuous across $r = r_0$ (Ref. 7). This situation affects some of the relations discussed above.

Points (a), (b), and (c) are obviously also valid in the case of anisotropic fluid, since they involve the continuity of e^λ and E^2 only. However, the results obtained under points (d) and (e) are in general no longer valid for anisotropic fluids.

(i) *Remark:* The radius r_0 of a charged anisotropic sphere can be less than the "classical electron radius" even when $W(r_0) > 0$ [see Theorem (d)].

In fact, in this case evaluating Eq. (2.15) at the boundary we obtain

$$M(1 - r_e/r_0) \geqslant (4\pi/3) p_\perp(r_0) r_0^3, \quad (2.23)$$

which indicates that when the tangential pressure is negative at the boundary certain anisotropic distributions can exist, showing gravitational repulsion ($r_0 \leqslant r_e$) even if $W(r_0) > 0$.

(ii) *Remark:* Unlike the perfect fluid case the interior of a charged anisotropic sphere with $r_0 \leqslant r_e$ is not necessarily gravitationally repulsive at every point when $W \leqslant 0$ throughout the body.

This becomes evident from Eq. (2.21) since if the term $(p_\perp - p_r)$ is negative then M_G can become positive inside the source even when $r_0 \leqslant r_e$ and $W \leqslant 0$.

(iii) *Remark:* Regarding Corollary (e2) it is possible to construct models of anisotropic charged sources with $r_0 = r_e$ that, contrary to the perfect fluid case, are gravitationally attractive at every point. We will show examples of this kind in Sec. III C. We will also see models of anisotropic spheres wherein the effective gravitational mass M_G is zero at all interior points. We also notice that like in the perfect fluid case the gravitational repulsion could also occur in the interior of anisotropic spheres with $r_0 > r_e$. In the present case this phenomenon is enhanced not only by a negative gravitational field energy W but also by positive values of $(p_\perp - p_r)$. We finish this section by mentioning that the interior of an anisotropic charged sphere can be gravitationally repulsive even when $W > 0$.

III. A NEW CLASS OF SOURCES OF REPULSIVE GRAVITATION

The discussion of the preceding section has elucidated questions (1)–(3) of the Introduction. In this section we are going to investigate questions (4) and (5). The search procedure we employ here consists in solving the field equations (2.3)–(2.6) to find examples (or counterexamples) elucidating these questions. The specific model we shall show is motivated by physical considerations and will also be useful to illustrate the results of Sec. II.

As we have seen, the purely gravitational field energy W plays a significant role in the size of the source. Motivated by this we will introduce this term explicitly into the field equations.

Using Eqs. (2.3) – (2.5), (2.14), and (2.16) we find

$$W = \frac{r}{6} - \frac{r^3 e^{-\lambda}}{6} \times \left[\frac{\nu''}{2} + \frac{\nu'^2}{4} + \frac{\lambda' - \nu'}{2r} - \frac{\nu'\lambda'}{4} + \frac{1}{r^2} \right]. \quad (3.1)$$

Subtracting Eq. (2.5) from (2.4) and using (3.1) we get

$$1 - e^{-\lambda} - r\lambda' e^{-\lambda}/2 = r^2\Delta + 3W/r, \quad (3.2)$$

where

$$\Delta \equiv 4\pi(p_\perp - p_r) + E^2. \quad (3.3)$$

From Eqs. (2.9) and (3.2) it follows that

$$r^3\Delta + 3W = -4\pi \int_0^r r^3 (T_0^0)' \, dr. \quad (3.4)$$

This expression does not contain the metric coefficients and relates the unknown $\rho, p_r, p_\perp, E^2$ to the contribution of the free gravitational field alone. We will use this equation below in constructing anisotropic charged models.

A. Generation of anisotropic models

To obtain specific models of perfect fluid sources one has to specify *a priori* two additional relations in such a way that the field equations become integrable. In the case of anisotropic matter due to the additional degree of freedom introduced, one needs to assume another additional relation. The ideal approach would be to know the relation between p_r and $p_\perp$ on physical grounds (an equation of state for the stresses). However, since this seems to be very difficult to carry out at present we shall instead construct anisotropic generalizations of perfect fluid charged models by using the following simplifying assumption: We shall disregard the possible effects of the anisotropy on the distribution of the nongravitational energy density T_0^0 and on the distribution of the energy W associated with the free gravitational field. That is, we assume that T_0^0 and W are the same in both isotropic and anisotropic charged models. This assumption is certainly true near the center, where $p_r \approx p_\perp$, and in general is justified in the case of small anisotropies. We note also that it follows from Eqs. (3.1)–(3.4) that our assumption implies that the anisotropy does not change the space-time geometry inside the source. In other words the functional dependence on the radial coordinate r of e^{ν} and e^{λ}, in our anisotropic charged models, will be the same as that of their isotropic counterparts. This means that in some sense our assumption is a generalization of a method used by various authors in the literature to obtain (neutral) anisotropic models from known (neutral) isotropic solutions.[9,10,20]

B. A specific model

To construct specific models of charged spheres one should specify, in some way, how the matter and charge are distributed throughout the body. The simplest choice is to assume the following.

(i) The rate of decrease of the energy density T_0^0 is linear in r; i.e.,

$$(T_0^0)' = -(5\alpha^2/4\pi)r, \quad \alpha^2 = \text{const} > 0. \tag{3.5}$$

(ii) The charge is uniformly distributed throughout the source, i.e.,

$$E = \text{const} \times r. \tag{3.6}$$

With these assumptions we will construct explicit solutions of the field equations.

The assumption that T_0^0 and W are not affected by the anisotropy implies that Δ has the same functional dependence on r in the isotropic and the anisotropic cases. Then it follows from assumption (ii) that we must set

$$\Delta = Kr^2, \quad K = \text{const}. \tag{3.7}$$

Substituting Eqs. (3.5) and (3.7) into (3.4) we find

$$W = [(\alpha^2 - K)/3]r^5. \tag{3.8}$$

Using this expression to integrate Eq. (3.2) we get

$$e^{-\lambda} = 1 + Cr^2 + \alpha^2 r^4, \tag{3.9}$$

where C is a constant of integration.

Substituting these equations into (3.1) and making the transformations

$$e^{\nu} \equiv Y^2; \quad e^{-\lambda} = Z, \quad x = r^2, \tag{3.10}$$

and

$$u = \int \frac{dx}{\sqrt{Z}}, \tag{3.11}$$

we obtain the equation

$$\frac{d^2Y}{du^2} + \frac{(\alpha^2 - 2K)}{4} Y = 0, \tag{3.12}$$

which has three solutions; viz.,

$$Y_{\text{I}} = A \sin \omega u + B \cos \omega u, \quad \omega \equiv (\alpha^2 - 2K)^{1/2}/2; \tag{3.13}$$

$$Y_{\text{II}} = A \sinh \omega u + B \cosh \omega u, \quad \omega \equiv (2K - \alpha^2)^{1/2}/2; \tag{3.14}$$

$$Y_{\text{III}} = A + Bu; \quad 2K = \alpha^2; \tag{3.15}$$

where A and B are constants of integration to be determined from the boundary conditions and u is obtained from Eqs. (3.9) and (3.11) as

$$u = \begin{cases} (1/|\alpha|)\ln(2|\alpha|\sqrt{1 + Cx + \alpha^2 x^2} + 2\alpha^2 x + C), \\ \quad \text{for } \alpha^2 > 0, \\ (2/C)\sqrt{1 + Cx}, \quad \text{for } \alpha = 0. \end{cases} \tag{3.16}$$

Substituting Eq. (3.9) into (2.3) we find

$$8\pi\rho = -3C - 5\alpha^2 r^2 - E^2, \tag{3.17}$$

from the continuity of e^{λ} across the boundary $r = r_0$ we get

$$C = -2M/r_0^3 + q^2/r_0^4 - \alpha^2 r_0^2. \tag{3.18}$$

Using this expression and that $E^2 = q^2 r^2/r_0^6$ (uniform charge density) we obtain the mass energy density ρ as follows:

$$8\pi r_0^2 \rho(r) = \frac{6M}{r_0}\left(1 - \frac{r_e}{2r_0}\right) + 3\alpha^2 r_0^4 - \left(5\alpha^2 r_0^4 + \frac{q^2}{r_0^2}\right)\left(\frac{r}{r_0}\right)^2. \tag{3.19}$$

Because the density decreases outward, its positiveness is assured by the requirement that ρ be positive at the boundary, viz.,

$$r_0 \geqslant \tfrac{2}{3} r_e [1 + \alpha^2/(2K - 8\pi p_\perp(r_0)/r_0^2)], \tag{3.20}$$

where

$$K - 4\pi p_\perp(r_0)/r_0^2 = q^2/r_0^6 > 0.$$

Equation (3.20) illustrates points (c)–(e) discussed in Sec. II. Specifically, this equation shows that when $W > 0$ $(\alpha^2 > K)$ the radius of the charged perfect fluid source is always larger than r_e (and consequently there is no gravitational repulsion around the body). If $\alpha^2 = K$, the space-time is conformally flat and the radius of the perfect fluid sphere is such that $r_0 \geqslant r_e$, where the equality holds for gaseous $[\rho(r_0) = 0]$ sources only. Nevertheless, when $W > 0$ and the source has anisotropic pressures, Eq. (3.20) shows that different from the perfect fluid case the radius of the body can be less than r_e if $p_\perp(r_0) < 0$. Moreover we see that, in principle, for large negative values of $p_\perp(r_0)$, r_0 can be as near as one wants to $2r_e/3$. We recall, however, that according to Eq. (2.13), this limiting value is only possible for gaseous distributions in the case $\alpha = 0$. When $W < 0$ $(\alpha^2 < K)$ the difference between isotropic and anisotropic matter, regarding the size of the source, is less drastic since the radii of both kinds of sources can be less than r_e.

Using the boundary conditions, namely, the continuity of e^{ν}, e^{λ}, $(e^{\nu})'$, and E^2 across the boundary, we obtain the final form of the solutions as follows.

Case I: If $\alpha^2 > 2K$ we find from Eqs. (3.10) and (3.13),

$$e_{\text{I}}^{\nu} = \{(\epsilon/2)(2\beta - 1)[\sin \omega(u - u_0)]/(\omega r_0^2) + Z_0^{1/2} \cos \omega(u - u_0)\}^2, \tag{3.21}$$

where

$$\begin{aligned} &\epsilon = M/r_0, \quad v = r/r_0, \quad \beta = (1 - r_e/2r_0), \\ &Z_0 = e^{-\lambda(r_0)} = 1 - 2M/r_0 + q^2/r_0^2 = 1 - 2\epsilon\beta, \end{aligned} \tag{3.22}$$

and $u_0 \equiv u(r_0)$ is obtained from Eq. (3.16). From Eq. (3.20) we obtain the range of β, namely, $1/4 \leqslant \beta \leqslant 1$. In the present case the Tolman–Whittaker effective mass, as given by Eq. (2.19), is

$$(M_G)_{\rm I} = Mv^3[(2\beta-1)\cos\omega(u-u_0) - (2Z_0^{1/2}/\epsilon)(\omega r_0^2)\sin\omega(u-u_0)]. \qquad (3.23)$$

The radial pressure is as follows:

$$8\pi r_0^2(p_r)_{\rm I} = \frac{\epsilon(2\beta-1)}{2} e_{\rm I}^{-\nu/2}\left\{(\alpha^2 r_0^4)(v^2-1)+2\epsilon v^2 - 2\epsilon\beta(1+v^2) - \frac{8(\omega^2 r_0^4)(Z_0 Z)^{1/2}}{\epsilon(2\beta-1)}\right\} \times\frac{\sin\omega(u-u_0)}{(\omega r_0^2)} + Z_0^{1/2} e_{\rm I}^{-\nu/2}\{2\epsilon[v^2 - (Z/Z_0)^{1/2}] + 2\epsilon\beta[2(Z/Z_0)^{1/2}-1-v^2] - \alpha^2 r_0^4(1-v^2)\}\cos\omega(u-u_0). \qquad (3.24)$$

Case II: If $\alpha^2 < 2K$ we find from Eqs. (3.10) and (3.14),

$$e_{\rm II}^{\nu} = \{(\epsilon/2)(2\beta-1)[\sinh\omega(u-u_0)]/(\omega r_0^2) + Z_0^{1/2}\cosh\omega(u-u_0)\}^2. \qquad (3.25)$$

The effective gravitational mass is

$$(M_G)_{\rm II} = Mv^3[(2\beta-1)\cosh\omega(u-u_0) + (2Z_0^{1/2}/\epsilon)(\omega r_0^2)\sinh\omega(u-u_0)]. \qquad (3.26)$$

The pressure can be obtained from Eqs. (3.24) by changing $\omega\rightarrow i\omega$ and $e_{\rm I}^{\nu}\rightarrow e_{\rm II}^{\nu}$.

Case III: If $2K=\alpha^2$ we obtain, from Eqs. (3.10) and (3.15),

$$e_{\rm III}^{\nu} = [Z_0^{1/2} + (\epsilon/2)(2\beta-1)(u-u_0)/r_0^2]^2. \qquad (3.27)$$

The effective mass is

$$(M_G)_{\rm III} = Mv^3(2\beta-1). \qquad (3.28)$$

The radial pressure may be obtained from Eq. (3.24) by taking the limit $\omega\rightarrow 0$.

In the three cases

$$e^{-\lambda} = Z = 1 - [2\epsilon\beta + (\alpha^2 r_0^4)]v^2 + (\alpha^2 r_0^4)v^4. \qquad (3.29)$$

The anisotropy is given by

$$4\pi r_0^2(p_\perp - p_r) = [(K r_0^4) - 2\epsilon(1-\beta)]v^2. \qquad (3.30)$$

The electric field is

$$r_0^2 E^2 = 2\epsilon(1-\beta)v^2 \qquad (3.31)$$

and the function W as well as the energy density are given by Eqs. (3.8) and (3.19), respectively.

C. Properties of the model

Examination of Eqs. (3.21)–(3.31) reveals that the above solutions are nonsingular and well behaved within some range of the dimensionless parameters $(\alpha^2 r_0^4)$, $(K r_0^4)$, ϵ, and β. Furthermore there are six different subcases, namely, the following.

(i) If $\alpha=0$, $K<0$, the solution is given by case I and $W>0$. The distribution does not contain the limiting case of charged perfect fluid. The radius of the anisotropic spheres can take all values $\geqslant 2r_e/3$.

(ii) If $\alpha=0$, $K>0$, the solution is given by case II and $W<0$. The distribution may have either isotropic or anisotropic pressures. In both situations the radius of the sphere is $\geqslant 2r_e/3$.

(iii) If $\alpha=0, K=0$, the solution is given by case III and the space-time is conformally flat ($W=0$). There is no charged perfect fluid and $r_0 \geqslant 2r_e/3$.

(iv) If $\alpha^2>0$, $K<\alpha^2/2$, the solution is given by case I and $W>0$. There are three different kinds of distributions depending on whether $K<0$, $K=0$, or $K>0$. For $K\leqslant 0$ the fluid can have anisotropic pressures only and the radius of the source is $>2r_e/3$. However, for $K>0$ the distribution may have either isotropic or anisotropic pressures. For perfect fluid $r_0>r_e$ and for anisotropic fluid $r_0>2r_e/3$.

(v) If $\alpha^2>0$, $K>\alpha^2/2$, the solution is given by case II, for $K\in(\alpha^2/2,\alpha^2)$, $W>0$ and the radius of the sphere is $>r_e$ when the fluid is perfect and $>2r_e/3$ when the fluid has anisotropic pressures. If $K=\alpha^2$, $W=0$ and unlike the subcase (iii) there are perfect fluid charged spheres, but with $r_0 \geqslant r_e$. There are also anisotropic spheres with $r_0>2r_e/3$. If $K>\alpha^2$, $W<0$ and for both perfect and anisotropic fluids $r_0>2r_e/3$.

(vi) If $\alpha^2>0$, $K=\alpha^2/2$, the solution is given by case III and $W>0$. For perfect fluid $r_0>r_e$ and for anisotropic fluid $r_0>2r_e/3$.

Subcases (i) and (ii), in the limit of the vanishing charge, reduce to anisotropic spheres of uniform density ($\rho=$ const) similar to those investigated by Bowers and Liang[7] and by Herrera and co-workers.[10] In the same limit, subcase (iii) becomes the Schwarzschild interior solution. In subcase (v) the solution with $K=\alpha^2$, in the limit of vanishing charge, is one of the solutions of Steward,[21] while the perfect fluid case with charge is one of the solutions of Shi-Chang.[22] This subcase (with $|\alpha| = 2\omega = C/2$) was also recently investigated by the present author[23] in connection with the "hoop conjecture."

For $\alpha\neq 0$ and for a perfect fluid our solutions reduce to some of those analyzed by Whitman and Burch.[24] Specifically, if we use their notation and put $C=-D$ and $\alpha^2=(k_0^2-\beta)$ we immediately recover six of their solutions, namely, those given by their Eqs. (5.3)–(5.9). Moreover their generalization of the Tolman IV solution as well as their charged Adler solution reduce to our solutions if we put $a=\beta=0$ in Eq. (5.11) and $D=0$ in Eq. (5.16a), respectively.

Finally if we set $|\alpha|=\omega$, $C=-16\pi^2\sigma_0^2 a^2/9$, $\alpha^2=(32\pi^2\sigma_0^2/45)$, and $K=16\pi^2\sigma_0^2/9$, then it can be verified by means of straightforward calculation that our subcase (v) (with $K>\alpha^2$ and $W<0$) reduces to the solution of Tiwari, Rao, and Kanakamedala. Thus we see that our solutions (3.21)–(3.31) can be considered as generalizations of some previously known solutions in the literature.

From the examination of Eqs. (3.21)–(3.31) we can see that although our models show gravitational repulsion, they, in general, neither satisfy the equation of state for false vacuum nor are their total gravitational masses M entirely of electromagnetic origin (i.e., M does not vanish when the charge vanishes). In this way we have constructed here a new class of sources of repulsive gravitation, which are different from those of TRK and of Gautreau. Thus our solutions elucidate questions (4) and (5) formulated in the Introduction.

Tables I and II show, respectively, the variation of the effective gravitational mass M_G, in the solutions I and II, for various values of r_0 (and fixed ϵ, α, K) in different regions of the source.

Some comments are in order.

First notice that Table I shows that the central region of the sources with $r_0 < r_e$ (in model I) is not gravitationally repulsive but attractive. The phenomenon of gravitational repulsion appears only near the boundary. This example gives information with regard to question (2) of the Introduction.

Second, Table II shows an example where the phenomenon of gravitational repulsion takes place in the interior of sources with $r_0 > r_e$. We see that M_G is negative inside $r < 0.8r_0$ when the radius of the source is 5/4 times the classical electron radius. This example illustrates our answer to question (3).

Third, notice that the charge makes a negative contribution to the effective gravitational mass inside the body. Specifically we see from the tables that for every value of $v = r/r_0$, M_G increases as the charge decreases. We will return to this point in Sec. IV.

Let us now consider the case where the radius of the source is equal to the classical electron radius. Setting $\beta = \frac{1}{2}(r_0 = r_e)$ in Eqs. (3.23), (3.26), and (3.28) we obtain

$$(M_G)_{\rm I} = -(2M/\epsilon)Z_0^{1/2}(\omega r_0^2)v^3 \sin\omega(u-u_0), \tag{3.32}$$

$$(M_G)_{\rm II} = (2M/\epsilon)Z_0^{1/2}(\omega r_0^2)v^3 \sinh\omega(u-u_0), \tag{3.33}$$

$$(M_G)_{\rm III} = 0. \tag{3.34}$$

We see that the effective gravitational mass M_G inside r_0 is zero in all three cases, as one expects. However, notice that within the body, i.e., for $r < r_0$, M_G may be positive, negative, or zero. This can be directly seen from Tables I and II and Eq. (3.34).

According to subcases (ii) and (v) the distribution corresponding to Eq. (3.33) may have both isotropic and anisotropic pressures. Moreover we find from Eq. (3.33) that $(M_G)_{\rm II} < 0$ throughout the body. The distributions corresponding the Eqs. (3.32) and (3.34) are more interesting, since according to subcases (i), (iii), (iv), and (vi) they can have anisotropic pressures only. Moreover from Eq. (3.32) we found that $(M_G)_{\rm I} > 0$ throughout the body. Thus we have here examples of distributions with $r_0 = r_e$ whose interior is gravitationally attractive. Equation (3.34) shows examples of distributions with $r_0 = r_e$ whose interior is gravitationally neutral (in the sense that the acceleration of gravity is zero inside the body), i.e., $(M_G)_{\rm III} = 0$ throughout. The latter two examples clearly illustrate the effects of the anisotropy on the phenomenon of gravitational repulsion discussed at the end of Sec. II. According to Corollary (e2) these examples have no perfect fluid analogs.

TABLE I. Variation of M_G/M with r in solution I given by Eqs. (3.21)–(3.23) for various values of r_0. We have taken $\epsilon = 0.4$, $Kr_0^4 = -0.3$, and $\alpha^2 r_0^4 = 0.1$.

r/r_0	$r_0 = 0.7r_e$	$r_0 = r_e$	$r_0 = 5r_e/4$	$r_0 = 5r_e/2$	$q = 0$
0	0	0	0	0	0
0.3	0.007	0.016	0.020	0.028	0.035
0.5	0.019	0.063	0.084	0.124	0.159
0.7	− 0.006	0.122	0.018	0.308	0.420
1	− 0.4	0	0.2	0.6	1

TABLE II. Variation of M_G/M with r in solution II given by Eqs. (3.25) and (3.26) for various values of r_0. We have taken $\epsilon = 0.4$, $Kr_0^4 = 0.3$, and $\alpha^2 r_0^4 = 0.1$.

r/r_0	$r_0 = 0.7r_e$	$r_0 = r_e$	$r_0 = 5r_e/4$	$r_0 = 5r_e/2$	$q = 0$
0	0	0	0	0	0
0.3	− 0.028	− 0.016	− 0.009	0.003	0.018
0.5	− 0.110	− 0.063	− 0.035	0.024	0.089
0.7	− 0.265	− 0.120	− 0.047	0.103	0.264
1	− 0.4	− 0	0.2	0.6	1

IV. SURFACE CHARGE

We now proceed to study the case of surface charge. We are here mainly interested in question (6) of the Introduction.

We will construct a specific solution using the same assumptions as in the preceding section [of course with the exception of assumption (ii)]. The interior solution is then given by Eqs. (3.9) and (3.13)–(3.15). In the present case to match the interior metric to the exterior Reissner–Nordström field across the charged boundary we apply the same method as used in Ref. 12. The final form of the solution is then obtained as follows:

$$e^{-\lambda} = Z = 1 - [2\epsilon\beta + (\alpha^2 r_0^4)]v^2 + (\alpha^2 r_0^4)v^4, \tag{4.1}$$

$$8\pi r_0^2\rho = 6\epsilon\beta + 3(\alpha^2 r_0^4) - 5(\alpha^2 r_0^4)v^2, \tag{4.2}$$

where ϵ, v, and β are still defined by Eq. (3.22). Since ρ decreases outward, its positiveness is now assured by the condition

$$r_0 \geqslant \tfrac{1}{2}r_e + \alpha^2 r_0^5/3\epsilon, \tag{4.3}$$

from which we obtain that in the present case $1 \leqslant \beta \leqslant 0$.

The anisotropy is given by

$$p_\perp - p_r = (Kr_0^2/4\pi)v^2. \tag{4.4}$$

The surface charge density σ is

$$\sigma^2 = [M(1-\beta)/8\pi^2 r_0^2]. \tag{4.5}$$

As in Sec. III, the calculation of e^ν leads to three different kinds of solutions, namely, the following cases.

Case I: If $\alpha^2 > 2K$,

$$e_{\rm I}^\nu = [(\epsilon\beta/2)[\sin\omega(u-u_0)]/\omega r_0^2 + Z_0^{1/2}\cos\omega(u-u_0)]^2, \tag{4.6}$$

$$M_{G_{\rm I}} = Mv^3[\beta\cos\omega(u-u_0) - (2(\omega r_0^2)/\epsilon)Z_0^{1/2}\sin\omega(u-u_0)]. \tag{4.7}$$

Case II: If $\alpha^2 < 2K$,

$$e_{\rm II}^\nu = [(\epsilon\beta/2)[\sinh\omega(u-u_0)]/\omega r_0^2 + Z_0^{1/2}\cosh\omega(u-u_0)]^2, \tag{4.8}$$

$$M_{G_{\rm II}} = Mv^3[\beta\cosh\omega(u-u_0)$$

$$+ (2(\omega r_0^2) Z_0^{1/2}/\epsilon) \sinh \omega(u - u_0)]. \qquad (4.9)$$

Case III: If $\alpha^2 = 2K$,

$$e^{\nu}_{\rm III} = [Z_0^{1/2} + (\epsilon\beta/2)(u - u_0)/r_0^2]^2, \qquad (4.10)$$

$$M_{G_{\rm III}} = Mv^3\beta. \qquad (4.11)$$

As we see the only mathematical difference between the above expressions and their counterparts of Sec. III resides in that the coefficient β, in Eqs. (4.6)–(4.11), replaces the term $(2\beta - 1)$ of Eqs. (3.21)–(3.28). Nevertheless, it is to be noted that this difference leads to effects of some consequences, namely, the following.

First, notice that in the case of charged spheres with $(T_0^0)' \leqslant 0$ the radius r_0 of the source is always $\geqslant 2r_e/3$. In the case of spheres with surface charge and $\rho' \leqslant 0$; however, r_0 can be as near to $r_e/2$ as one wants.

Second, the inspection of Eqs. (4.1)–(4.11) shows that the radius r_0 of a perfect fluid sphere with surface charge can be less than r_e independently of whether W is positive, negative, or zero. It is worthwhile to recall that in the case of perfect fluid charged spheres W necessarily becomes negative inside the body when $r_0 \leqslant r_e$.

The third point to notice is that the effective gravitational mass M_G inside the source is less in the case of volume charge than in the case of surface concentration of charge, for every specific value of r, ϵ, K, and α. To illustrate this fact we give in Tables III and IV the variation of M_G with r (with the same values of ϵ, α, and K as in Tables I and II) for the solutions described by Eqs. (4.6), (4.7) and (4.8), (4.9), respectively. The direct comparison of Table I with Table III and of Table II with Table IV clearly shows that the charge makes a negative contribution to M_G. This contribution can be easily calculated in the case of the solutions (4.6)–(4.11) by evaluating the discontinuity of the effective gravitational mass M_G at the boundary $r = r_0$ of the body. In fact, from Eqs. (4.7), (4.9), and (4.11) we find

$$M_G^-(r_0) = M\beta = M(1 - r_e/2r_0) \equiv M_G(r_0)|_{\rm interior} \qquad (4.12)$$

for the three cases. The exterior value of M_G is obtained from Eq. (1.1) as follows:

$$M_G^+(r_0) = M(1 - r_e/r_0) \equiv M_G(r_0)|_{\rm exterior}. \qquad (4.13)$$

Then the contribution of the surface charge to the effective gravitational mass of the body is given by

$$\Delta M_G \equiv M_G^+(r_0) - M_G^-(r_0) = -\frac{q^2}{2r_0} = -\frac{M}{2}\frac{r_e}{r_0}; \qquad (4.14)$$

as we see this contribution is negative for all values of r_0, implying that the shell of charge is always gravitationally repulsive.

Fourth, notice that in cases I and III given by Eqs. (4.6), (4.7) and (4.10), (4.11), the effective gravitational mass M_G is always positive at all points within the body for all values of r_0. In other words when the radius of the source is $< r_e$ there is gravitational repulsion only in the vicinity of the body. Inside the body there is no gravitational repulsion but attraction. We recall that this is not so in the case of charged spheres wherein M_G necessarily becomes negative inside the body when $r_0 < r_e$.

Table IV shows that in solution II, given by Eqs. (4.8), (4.9), M_G is negative in the central region of the body for $r_0 \leqslant r_e$. This solution therefore indicates that even in the case where the charge is concentrated entirely at the boundary, the interior of the body may become gravitationally repulsive.

Until now the only known solution in the literature representing a charged spherical shell whose interior is gravitationally repulsive has been the solution of Cohen and Cohen[6] (CC); therefore it may be worthwhile to emphasize the differences between our example (case II) and CC's: (a) in CC's solution the source is gravitationally repulsive at every interior point independently of its radius r_0, in our case II the phenomenon appears only in the central region of the body when $r_0 \leqslant r_e$; and (b) in CC's solution the interior is described by the static form of the de Sitter metric so that $\rho = -p = \text{const}$, moreover all physical quantities vanish in the limiting case of vanishing charge, whereas in our solution II, $\rho + p \neq 0$ and when the surface charge vanishes we obtain a neutral distribution with $M > 0$.

We close this section by noticing that in the case of perfect fluid our solution III with $\alpha = 0$ reduces to the Schwarzschild interior solution with surface concentration of charge. This solution has been discussed by Stettner[25] and by Whitman and Burch.[24] Finally for $\rho = 0$ our models reduce to a bubble having a charged surface of radius $r_0 = r_e/2$, in agreement with Eq. (2.12). Inside the surface the spacetime is flat so that M_G is negative for $r_0 < r < 2r_0$ and zero for $r < r_0$.

V. SUMMARY AND CONCLUSIONS

In this work we have studied the phenomenon of gravitational repulsion in static sources of the Reissner–Nordström field. We have considered three different kinds of

TABLE III. Variation of M_G/M with r in solution I with surface charge given by Eqs. (4.6) and (4.7) for various values of r_0. In this solution $r_0 > r_e/2$. To facilitate the comparison with Table I we have again taken $\epsilon = 0.4$, $Kr_0^4 = -0.3$, and $\alpha^2 r_0^4 = 0.1$.

r/r_0	$r_0 = 0.51r_e$	$r_0 = 0.7r_e$	$r_0 = r_e$	$r_0 = 5r_e/4$	$r_0 = 5r_e/2$	$q = 0$
0	0	0	0	0	0	0
0.3	0.010	0.024	0.028	0.030	0.033	0.035
0.5	0.071	0.100	0.120	0.131	0.140	0.159
0.7	0.130	0.220	0.280	0.310	0.370	0.420
1	0.01	0.3	0.5	0.6	0.8	1

TABLE IV. Variation of M_G/M with r in solution II with surface charge given by Eqs. (4.8) and (4.9) for various values of r_0. In this solution $r_0 > r_e/2$. To facilitate the comparison with Table II we have again taken $\epsilon = 0.4$, $Kr_0^4 = 0.3$, and $\alpha^2 r_0^4 = 0.1$.

r/r_0	$r_0 = 0.51r_e$	$r_0 = 0.7r_e$	$r_0 = r_e$	$r_0 = 5r_e/4$	$r_0 = 5r_e/2$	$q = 0$
0	0	0	0	0	0	0
0.3	− 0.010	− 0.008	− 0.001	0.001	0.009	0.018
0.5	− 0.060	− 0.027	0.002	0.018	0.051	0.089
0.7	− 0.120	− 0.010	0.055	0.094	0.174	0.264
1	0.01	0.3	0.5	0.6	0.8	1

sources, namely, perfect fluid charged spheres, anisotropic charged spheres, and spheres whose net charge is concentrated at the boundary surface alone.

We have found that in the case of charged perfect fluid spheres there is a close connection between the sign of the purely gravitational field energy $W = rC^3{}_{232}/2$ and the phenomenon of gravitational repulsion. Specifically, regarding question (1) we obtained that the necessary condition for $r_0 \leqslant r_e$ is that $W(r_0) < 0$. For what concerns question (2) we showed that a sphere with $r_0 \leqslant r_e$ is gravitationally repulsive at all interior points when $W \leqslant 0$ throughout the body.

Unfortunately these relations are no longer valid neither in the case of charged anisotropic matter nor if the charge is concentrated at the surface of the body. Returning to question (2), we have seen that there can exist situations where the effective gravitational mass M_G is positive in the central region of the body even when its radius is less than r_e. Moreover, the results of Sec. IV show that in the case of surface charge, M_G may be positive throughout the body for all values of r_0. From this we can conclude that the only possible case where a source of the Reissner–Nordström field could give rise to gravitational repulsion, in the vicinity of the body and at the same time be gravitationally attractive at all interior points, is in the case where the net charge resides entirely on the boundary surface.

The answer to question (3) is that certain bodies can exist showing gravitational repulsion even if their radius r_0 is somewhat larger than r_e. Of course the region where $M_G < 0$ must be situated in the interior of the body. In this sense such distributions are similar to those recently discussed by the present author.[3]

Questions (4) and (5) were discussed on the basis of a specific family of solutions of the Einstein–Maxwell equations. This family was constructed by using physical considerations and includes some known solutions from the literature. Our answer to question (4) is that there are solutions representing sources of repulsive gravitation in which the equation of states of false vacuum $\rho = -p$ does not apply. In these models M_G becomes negative inside the source by the negative contribution due to the charge rather than by the polarization of vacuum.[1] Question (5) has a similar answer. We found that in general the total mass M (measured by an observer at infinity) of a source of repulsive gravitation does not vanish when the charge vanishes, rather, in this limit we obtain neutral distributions continuously matched to the exterior Schwarzschild solution with $M > 0$. This means that in general only certain amount, but not all, of the mass of a source of repulsive gravitation is of electromagnetic origin.

Question (6) was discussed in Sec. IV. We have seen that although the only difference between systems with surface charge and their counterparts with volume charge resides in the evaluation of the constants of integration at the boundary, there are some difference. One of the most interesting differences is that the radius of the body, with positive and decreasing outward energy density, is $\geqslant r_e/2$ for surface charge and $\geqslant 2r_e/3$ for volume charge. We also discussed that the charge makes a negative contribution to the effective gravitational mass M_G. This effect occurs even when the charge resides entirely on the surface of the body. Specifically, an increase in the surface charge causes a decrease in M_G inside the charged shell. Therefore, in principle, the interior of such sources can become gravitationally repulsive. In the explicit solutions we constructed in Sec. IV there are examples illustrating this latter point, as it is seen in Table IV.

The introduction of anisotropy allowed us to find two new kinds of distributions representing charged spheres with radius equal to the classical electron radius, namely, Eqs. (3.32) and (3.34). In one of these the effective gravitational mass M_G inside the body is positive while in the other it is zero. Therefore as was discussed at the end of Sec. III, the interior of such bodies is not gravitationally repulsive, contrary to what would happen in a perfect fluid sphere.

The above results have been obtained subject only to the conditions that all physical quantities be finite everywhere, T_0^0 and ρ be non-negative and decreasing outward and $p_\perp = p_r$ at the origin. The gravitational repulsion in the models results from a violation of the "strong energy condition" $[(T_{\mu\nu} - g_{u\nu}T/2)U^\mu U^\nu > 0$, where U^μ is any timelike vector$]$ which basically says that gravitation is always an attractive force. The violation of this condition, as well as of the "weak energy condition," has been discussed by a number of authors[26–33] in order to avoid the singularities predicted by the Hawking–Penrose theorem (Ref. 16, p. 266).

If we assume that the strong energy condition can never be violated then it follows from the results of this work that the minimum radius of any charged sphere in equilibrium is just the classical electron radius, i.e.,

$$r_0 = r_e = q^2/M. \qquad (5.1)$$

Moreover this minimal size would be attainable only by bodies with anisotropic pressures (since for perfect fluid M_G becomes negative inside the sphere when $r_0 = r_e$).

To finalize and in order to avoid possible ambiguities in the concept of gravitational repulsion outside the sources of the Reissner–Nordström field,[34,35] we briefly discuss the radial motion of neutral test particles in this field. Using Eqs.

(2.8) to integrate the equation of geodesic (with θ and ϕ constants) we find

$$V^2 = 1 - (m_0/\epsilon_0)^2 f^2, \tag{5.2}$$

$$\frac{dV}{d\tau} = -\frac{M}{r^2}\left(1 - \frac{r_e}{r}\right)(1 - V^2) f^{-1}, \tag{5.3}$$

where $d\tau$ is the locally measured time, viz., $d\tau = f\,dt$, V is the locally measured radial velocity, ϵ_0 is the (constant) energy of the test particle, and m_0 its mass. To a distant observer whose meter sticks and clocks are not affected by gravity the radial velocity and acceleration are given by

$$\frac{dr}{dt} = Vf^2, \tag{5.4}$$

$$\frac{d^2 r}{dt^2} = \frac{M}{r^2}\left(1 - \frac{r_e}{r}\right)(3V^2 - 1) f^2. \tag{5.5}$$

It follows from Eq. (5.3) that a local observer finds gravitational repulsion (for all values of $V^2 < 1$) in the region $r < r_e$ only. However, from Eq. (5.5) we see that to a distant observer the gravitational repulsion can occur in the region $r > r_e$. In fact, to a distant observer a neutral test particle with $V > 1/\sqrt{3}$ will be repelled in $r > r_e$ and attracted in $r < r_e$. For $V < 1/\sqrt{3}$ the field is attractive in $r > r_e$ and repulsive in $r < r_e$ for both observers. This effect is similar to that discussed in an interesting paper by McGruder[36,37] for the Schwarzschild field and can be attributed to the fact that the time and space (radial) intervals measured by a distant observer differ from the time and space (radial) intervals measured by a local observer.

[1] Ø. Grøn, Phys. Rev. D **31**, 2129 (1985).

[2] R. Gautreau, Phys. Rev. D **31**, 1860 (1985).

[3] J. Ponce de Leon, J. Math. Phys. **28**, 410 (1987).

[4] J. M. Cohen and R. Gautreau, Phys. Rev. D **19**, 2273 (1979).

[5] R. N. Tiwari, J. R. Rao, and R. R. Kanakamedala, Phys. Rev. D **30**, 489 (1984).

[6] J. M. Cohen and M. D. Cohen, Nuovo Cimento **60**, 241 (1969).

[7] R. L. Bowers and E. P. T. Liang, Astrophys. J. **188**, 657 (1974).

[8] P. S. Letelier, Phys. Rev. D **22**, 807 (1980).

[9] S. S. Bayin, Phys. Rev. D **26**, 1262 (1982).

[10] M. Consenza, L. Herrera, M. Esculpi, and L. Witten, J. Math. Phys. **22**, 118 (1981).

[11] J. Ponce de Leon, J. Math. Phys. **28**, 1114 (1987).

[12] L. Herrera and J. Ponce de Leon, J. Math. Phys. **26**, 2302 (1985).

[13] W. B. Bonnor, Phys. Lett. A **99**, 424 (1983).

[14] K. P. Tod, Proc. R. Soc. London, Ser. A **388**, 462 (1983).

[15] It follows from Eq. (2.13) that the total mass M of a body of radius r_0 and charge q satisfies the relation $M > 4M^{(1)}/3$, where $M^{(1)}$ is the electromagnetic mass corresponding to the case where the charge is uniformly distributed in the shape of a spherical shell of radius r_0, viz., $M^{(1)} = q^2/2r_0$. It is interesting to note that the factor $\frac{4}{3}$ also appears in connection with the construction of classical models of the electron. See for instance, J. Schwinger, Found. Phys. **13**, 373 (1983).

[16] S. W. Hawking and G. F. R. Ellis, *The Large Scale Structure of Spacetime* (Cambridge U.P., London, 1973), p. 85.

[17] M. E. Cahill and G. C. McVittie, J. Math. Phys. **11**, 1383 (1970).

[18] E. N. Glass, J. Math. Phys. **20**, 1508 (1979).

[19] J. Krishna Rao, Gen. Relativ. Gravit. **4**, 5 (1973).

[20] J. Carot and J. Ibanez, J. Math. Phys. **26**, 2282 (1985).

[21] B. W. Steward, J. Phys. A **15**, 2419 (1982).

[22] Z. Shi-Chang, Gen. Relativ. Gravit. **15**, 293 (1983).

[23] J. Ponce de Leon, Gen. Relativ. Gravit. **19**, 289 (1987).

[24] P. G. Whitman and R. C. Burch, Phys. Rev. D **24**, 2049 (1981).

[25] R. Stettner, Ann. Phys. (NY) **80**, 212 (1973).

[26] F. J. Tipler, Phys. Rev. D **17**, 2521 (1978).

[27] T. A. Roman and P. G. Bergman, Phys. Rev. D **28**, 1265 (1982).

[28] J. D. Bekenstein, Phys. Rev. D **11**, 2521 (1975).

[29] G. L. Murphy, Phys. Rev. D **8**, 4231 (1973).

[30] P. N. Baaklini, Phys. Lett. A **66**, 357 (1978).

[31] L. Parker and S. A. Fulling, Phys. Rev. D **7**, 2357 (1973).

[32] Yu. A. Beletsky, Preprint ITP-83-148E of Academy of Sciences of The Ukrainian S.S.R., Institute for Theoretical Physics, 1983.

[33] Yu. A. Beletsky, Preprint ITF-84-7P of Academy of Sciences of The Ukrainian S.S.R., Institute for Theoretical Physics, 1984 (in Russian).

[34] A. Qadir, Phys. Lett. A **99**, 419 (1983), and references therein.

[35] Ø. Grøn, Phys. Lett. A **94**, 424 (1983), and references therein.

[36] C. H. McGruder III, Phys. Rev. D **25**, 3191 (1982).

[37] The gravitational repulsion outside The Reissner–Nordström field has recently been discussed by K. D. Krori and M. Barua, Phys. Rev. D **31**, 3135 (1985); however, this effect has not been explicitly displayed in their work.

Stochastic calculus on distorted Brownian motion

Dražen Pantić
Laboratory 180/02, Boris Kidrič Institute, P.O.B. 522, 11000 Belgrade, Yugoslavia

(Received 22 August 1986; accepted for publication 15 July 1987)

Relations between Malliavin's calculus and the formulation of Schrödinger dynamics in terms of local Dirichlet forms are investigated. It is proved that under mild regularity conditions there exists a directional derivative of the solution of the stochastic equation associated with the corresponding Dirichlet form. Also, sufficient conditions are given so that $E\{f(X_{t,x})\}$ is a strong solution of the Cauchy problem $(\partial/\partial t)\, h(t,x) = (-\Delta + Q(x))h(t,x)$, where $\{X_{t,x}, t>0\}$, $x\in\mathbb{R}^d$, is a distorted Brownian motion.

I. INTRODUCTION

In this paper we shall follow the formulation of Schrödinger dynamics in terms of local Dirichlet forms and the corresponding Markov process[1] and relate this approach to Malliavin's calculus on Brownian motion.

Let $H = -\Delta + Q$ be a Schrödinger operator H: $\mathrm{dom}(H) \to L_2(l)$ (where l is the Lebesque measure). The basic assumption is that there exists a "ground state" g_0, i.e.,

$$(\exists g_0 \in L_2(l))(Hg_0 = 0 \text{ and } g>0 \text{ almost sure } (l)). \tag{1}$$

If we define measure m as

$$m(dx) = g_0^2(x) l(dx), \tag{2}$$

then it is obvious that

$$f\in L_2(m) \Leftrightarrow fg_0 \in L_2(l), \tag{3}$$

and for $f_i = g_0 f_i$, $i = 1,2$, where f_i is a twice continuously differentiable function with compact support, the matrix element could be computed[1] as

$$(f_1|Hf_2)_{L_2(l)} = (\nabla f_1|\nabla f_2)_{L_2(m)}, \tag{4}$$

where $(\ |\ )_W$ is a scalar product in Hilbert space W.

So, relation (4) defines a quadratic form

$$\epsilon(f,g) = (\nabla f|\nabla g)_{L_2(m)}, \tag{5}$$

$\mathrm{Dom}(\epsilon) = \mathbb{C}_0^1$. The form ϵ is a densely defined symmetric positive form. If $g_0 \in L_{2,\mathrm{loc}}(\mathbb{R}\backslash N)$, where N is a closed set, $l(N) = 0$, then the form is closable,[2] and there exists a self-adjoint operator $\widetilde{H}$, $\mathrm{Dom}(\widetilde{H}) \subseteq \mathrm{Dom}(\epsilon)$ so that $f\in\mathrm{Dom}(\epsilon)$,

$$g\in \mathrm{Dom}(\widetilde{H}) \Rightarrow \epsilon(f,g) = (f|\widetilde{H}g)_{L_2(m)}, \tag{6}$$

and, if we define

$$b = 2\nabla g_0/g_0, \tag{7}$$

for f a twice continuously differentiable function, then

$$\widetilde{H}f = (\Delta f + b\cdot\nabla f). \tag{8}$$

The connection between the form ϵ and the Markov processes is given by the next theorem.[3]

Theorem 1.1: If the form ϵ defined in (5) is closable, then there exists a diffusion process $\{X_t, t>0\}$ with values in $\mathbb{R}^d$, and transition semigroup $\{P_f, t>0\}$ symmetric in $L_2(m)$, such that $P_t = \exp(-t\widetilde{H})$. □

As a consequence it is obvious that m is an invariant measure for the process $\{X_t\}$.[1]

The next theorem (the existence theorem[2]) links the process X and Brownian motion.[2]

Theorem 1.2: Let $\{X_t, t>0\}$ be a diffusion process associated with the form ϵ as in Theorem 1.1. If $g_0\in L_2(l)$, $(\nabla g_0)_i \in L_2(l)$, $i = 1,d$, then

$$X_t - X_0 = \int_0^t b(X_s)ds + B_t, \tag{9}$$

where b is defined in (7) and $\{B_t, t>0\}$ is a standard d-dimensional Brownian motion, starting from 0. □

Following Ref. 4 we shall call the process X "distorted Brownian motion."

If we define $\tilde{h}(t,x) = P_t f(x), f\in L_2(m)$, then $\tilde{h}$ is a weak $L_2(m)$ solution of the Cauchy problem

$$\frac{\partial}{\partial t}\tilde{h}(t,x) = \widetilde{H}\tilde{h}(t,x), \quad \tilde{h}(0,x) = f(x) \tag{10}$$

(see Ref. 2), and the function $h(t,x) = g_0(x)\tilde{h}(t,x)$ is a weak $L_2(l)$ solution of the problem

$$\frac{\partial}{\partial t}h(t,x) = Hh(t,x), \quad h(0,x) = f(x)g_0(x).$$

Now we shall use Bismut's approach to Malliavin's calculus on Brownian motion and give the sufficient condition so that the function $P_t f$ will be the strong solution of Eq. (10).

II. STOCHASTIC CALCULUS

In this paper we shall state all the results for the one-dimensional case. The multidimensional extension is straightforward.

Let us mark with (A) the condition that the function b has a continuous bounded derivative, and that $b\in L_2(m)$.

Under condition (A) there exists a unique solution $\{X_{t,x}, t\geqslant 0\}$ of Eq. (9) with $X_{0,x} = x$.[5] Using the standard method of successive approximation, we can conclude that for each $t\geqslant 0$, $X_{t,x}$ is a continuous function in $x\in\mathbb{R}$.

Let Π denote a class of mesurable processes adapted to the family of σ-algebras $\sigma\{B_s\ s<t\}$, $t>0$, so that for $t>0$, $r>0$, $u\in\Pi$,

$$E\left\{\exp\left(r\int_0^t u_s^2\, ds\right)\right\} < \infty.$$

The first lemma asserts that condition (A) implies the existence of the directional derivative of $X_{t,x}$ in the sense of Bismut.[6]

Lemma 1: If (A) is fulfilled and u is of the class Π, then there exists a unique solution of

0022-2488/88/010207-03$02.50

$$X^{r,u}_{t,x} - x = \int_0^t b(X^{r,u}_{s,x})ds + B_t + r\int_0^t u_s\, ds, \qquad (11)$$

and the directional derivative in the direction u,

$$D_u X_{t,x} = \lim_{r\to 0}(X^{r,u}_{t,x} - X_{t,x})/r$$

exists almost sure (a.s.), and

$$D_u X_{t,x} = \exp\left(\int_0^t b'(X_{v,x})dv\right) \times \int_0^t u_s \exp\left(\int_0^s b'(X_{v,x})dv\right)ds. \qquad (12)$$

□

The proof is given in the Appendix.

Lemma 2: If the function b satisfies condition (A) then there exists the limit $DX_{t,x} = \lim_{y\to x}(X_{t,x} - X_{t,y})/(x-y)$ and

$$DX_{t,x} = \exp\left(\int_0^t b'(X_{v,x})dv\right) \qquad (13)$$

for $t \geqslant 0$. The proof is given in the Appendix. □

Now, from relations (12) and (13) it is obvious that

$$D_u X_{t,x} = DX_{t,x} \int_0^t \frac{u_s}{DX_{s,x}}\, ds, \qquad (14)$$

and, as the process $\{D_{t,x},\ t \geqslant 0\}$ belongs to the class Π, for $\tilde{u}_\cdot = DX_{\cdot,x}$, we get

$$D_{\tilde{u}} X_{t,x} = t\cdot DX_{t,x}, \quad t \geqslant 0. \qquad (15)$$

Next, if $f\in C^1_{b}$, then for $t > 0$ it follows that

$$f(X_{t,x}) - f(X_{t,y}) = \int_y^x f'(X_{t,z})DX_{t,z}\, dz = \int_y^x \frac{f'(X_{t,z})D_{\tilde{u}}X_{t,z}dz}{t}.$$

Hence it is obvious that

$$f(X_{t,x}) - f(X_{t,y}) = \int_y^x \frac{D_{\tilde{u}}(f(X_{t,z}))dz}{t}. \qquad (16)$$

Theorem 2.1: Under condition (A) the function $x \to P_t(x)$ is a continuously differentiable in $x\in\mathbb{R}$, for $t > 0$ and f a measurable bounded function.

The proof is given in the Appendix. □

Now we shall strengthen condition (A) assuming that the first and the second derivatives of b are continuous bounded functions, and that b belongs to $L_2(m)$. We shall mark this condition with (B).

An immediate consequence is that there exists a limit

$$DDX_{t,x} = \lim_{y\to x}(DX_{t,x} - DX_{t,y})/(x-y)$$

and (17)

$$DDX_{t,x} = DX_{t,x}\int_0^t b''(X_{v,x})DX_{v,x}\, dv.$$

Theorem 2.2: Let the function b satisfy condition (B). If f is a bounded measurable function then for $t > 0$, the mapping $x \to P_t f(x)$ is twice continuously differentiable in $\mathbb{R}$. The proof is given in the Appendix. □

From Theorem 2.2 it follows that for $f\in\mathbf{M}_{b}$, $P_t f\in\mathrm{Dom}(\widetilde{H})$, i.e., P_t: $\mathbf{M}_{b} \to \mathrm{Dom}(\widetilde{H})$. So $\{P_t,\ t > 0\}$ is a differentiable semigroup[7] on $\mathbf{M}_{b}$ and the function $P_t f$ is a strong solution of Eq. (10).

APPENDIX: PROOFS OF THE RESULTS

Proof of Lemma 1: The existence and uniqueness of the solution of Eq. (11) follows from Ref. 8, and the solution $\{X^{r,u}_{t,x},\ t \geqslant 0\}$ is continuous in $(t,r)\in\mathbb{R}\times\mathbb{R}$. Then, $X^{0,u}_{t,x} = X_{t,x}$, and

$$(X^{r,u}_{t,x} - X_{t,x}) = \int_0^t (b(X^{r,u}_{s,x}) - b(X_{s,x}))ds + r\int_0^t u_s\, ds$$
$$= \int_0^t b'(Y_{r,s})(X^{r,u}_{s,x} - X_{s,x})ds + r\int_0^t u_s ds,$$

where

$$\min(X^{r,u}_{s,x}, X_{s,x}) \leqslant Y_{s,r} \leqslant \max(X^{r,u}_{s,x}, X_{s,x}), \qquad (\mathrm{A}1)$$

so it is obvious that

$$\frac{(X^{r,u}_{t,x} - X_{t,x})}{r} = \exp\left(\int_0^t b'(Y_{s,r})ds\right)\int_0^t u_s \exp\left(\int_0^s b'(Y_{v,r})dv\right)ds.$$

As $Y_{s,r} \to X_{s,x}$ $(r\to 0)$, and b' is a bounded continuous function, it follows that $D_u X_{t,x}$ exists and relation (12) holds. Q.E.D

The proof of Lemma 2 is basically the same as the proof of Lemma 1 so we shall not repeat it.

Proof of Theorem 2.1: Let $f\in\mathbf{M}_{b}$. Then from (16), by Fubini's theorem it follows that

$$P_t f(x) - P_t f(y) = \int_y^x E\{D_{\tilde{u}}(f(X_{t,z}))\}\frac{dz}{t}. \qquad (\mathrm{A}2)$$

As $f(X_{t,x})$ is a square integrable functional of Brownian motion, we can apply the integration by parts formula,[6] which asserts that for $u\in\Pi$,

$$E\{D_u(f(X_{t,z}))\} = E\left\{f(X_{t,z})\int_0^t u_s\, dB_s\right\},$$

so (A2) becomes

$$P_t f(x) - P_t f(y) = \int_y^x E\left\{f(X_{t,z})\int_0^t DX_{s,z}\, dB_s\right\}\frac{dz}{t}. \qquad (\mathrm{A}3)$$

Both sides of (A3) are linear functionals of f, so by using the limit process we can conclude that (A3) is valid for $f\in\mathbf{M}_{b}$. Hence the function $P_t f$ is absolutely continuous and by the semigroup property $P_{t+s} = P_t P_s f$,

$$P_{t+s}f(x) - P_{t+s}f(y) = \int_y^x E\left\{P_s f(X_{t,x})\int_0^t DX_{v,x}dB_v\right\}\frac{dz}{t}. \qquad (\mathrm{A}4)$$

Then, it follows that there exists

$$\lim_{y\to x}(P_{t+s}f(x) - P_{t+s}f(y))/(x-y)$$

and

$$P_t f(x) = E\left\{f(X_{t,x})\int_0^t DX_{v,x}dB_v\right\}t^{-1}. \qquad (\mathrm{A}5)$$

Q.E.D.

Proof of Theorem 2.2: Let $f \in \mathrm{C}^{1\,'}_{\ell}$. Then, from Theorem 2.1, it follows that $P_t f$ is continuously differentiable and if we define $Q_t f(x) = t(\partial/\partial x)P_t f(x)$, then

$$\begin{aligned} Q_t f(x) - Q_t f(y) &= E\left\{(f(X_{t,x}) - f(X_{t,y}))\int_0^t DX_{v,x}\, dB_v\right\} \\ &\quad + E\left\{f(X_{t,y})\int_0^t (DX_{v,x} - DX_{v,y})dB_v\right\}. \end{aligned}$$

Hence

$$\begin{aligned} \frac{\partial}{\partial x} Q_t f(x) &= E\left\{f'(X_{t,x})DX_{t,x}\int_0^t DX_{v,x}\, dB_v\right\} \\ &\quad + E\left\{f(X_{t,x})\int_0^t DDX_{v,x}\, dB_v\right\} \qquad (A6) \end{aligned}$$

[$DDX_{t,x}$ is defined in (17)].

As $\{DX_{t,x},\ t \geqslant 0\}$ is a process of locally bounded variation, then

$$\int_0^t DX_{v,x}\, dB_v = (DX_{t,x})B_t - \int_0^t B_v b'(X_{v,x})DX_{v,x}\, dv \qquad (A7)$$

and

$$\begin{aligned} &DX_{t,x}\int_0^t DX_{v,x}\, dB_v \\ &\quad = (DX_{t,x})^2 B_t - DX_{t,x}\int_0^t B_v b'(X_{v,x})DX_{v,x}\, dv. \qquad (A8) \end{aligned}$$

From relations (14) and (15), and the fact that

$$(DX_{t,x})^2 = DX_{t,x}\int_0^t b'(X_{v,x})DX_{v,x}\, dv + DX_{t,x},$$

it follows that

$$(DX_{t,x})^2 = D_{u_1}X_{t,x} + D_{\tilde u}X_{t,x}/t, \qquad (A9)$$

where

$$\tilde u_s = DX_{s,x}, \quad \text{and} \quad u_{1_s} = b'(X_{s,x})(DX_{s,x})^2, \quad s \geqslant 0. \qquad (A10)$$

On the other hand, Lemma 1 implies that

$$DX_{t,x}\int_0^t B_v b'(X_{v,x})DX_{v,x}\, dv = D_{u_2}X_{t,x}, \qquad (A11)$$

where

$$u_{2_s} = B_s b'(X_{s,x})(DX_{s,x})^2, \quad s \geqslant 0. \qquad (A12)$$

Then, as processes $\tilde u$, u_1, u_2 belong to the class Π formula, (A6) reads as

$$\begin{aligned} \frac{\partial}{\partial x} Q_t f(x) &= E\left\{f'(X_{t,x})\left[D_{u_1}X_{t,x} + \frac{D_{\tilde u}X_{t,x}}{t}\right]B_t \right. \\ &\quad \left. - f'(X_{t,x})D_{u_2}X_{t,x}\right\} \\ &\quad + E\left\{f(X_{t,x})\int_0^t DDX_{v,x}\, dB_v\right\}. \qquad (A13) \end{aligned}$$

Using the integration by parts formula[6] and keeping in mind that

$$D_u(G_1G_2) = G_1D_uG_2 + G_2D_uG_1$$

and

$$D_u(B_t) = \int_0^t u_v\, dv,$$

formula (A13) becomes

$$\begin{aligned} t\frac{\partial^2}{\partial x^2}P_t f(x) &= E\left\{f(X_{t,x})\left[\left(B_t\int_0^t u_{1_v}\, dB_v - \int_0^t u_{1_v}\, dv\right)\right.\right. \\ &\quad + \left(B_t\int_0^t \tilde u_v\, dB_v - \int_0^t \tilde u_v\, dv\right)(t)^{-1} \\ &\quad \left.\left. - \int_0^t u_{2_v}\, dB_v + \int_0^t DDX_{v,x}\, dB_v\right]\right\}. \qquad (A14) \end{aligned}$$

It could be easily shown that the process in the square brackets is a square integrable process. With similar reasoning as in the proof of Theorem 2.1 it could be concluded that formula (A13) is valid for $f \in \mathrm{M}_{\ell}$. Q.E.D

[1] L. Streit, Phys. Rep. **3**, 77 (1981).

[2] S. Albeverio, R. Høegh-Krohn, and L. Streit, J. Math. Phys. **18**, 907 (1977).

[3] M. Fukushima, *Dirichlet Forms and Markov Processes* (North-Holland, Amsterdam, 1980).

[4] H. Ezawa, J. R. Klauder, and L. A. Shepp, Ann. Phys. (NY) **88**, 588 (1974).

[5] A. Ju. Veretennikov, Usp. Mat. Nauk. **38**, 113 (1983).

[6] M. Zakai, Acta Appl. Math. **2**, 3, 175 (1985).

[7] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations* (Springer, Berlin, 1983).

[8] Yu. N. Blagovescenski and M. I. Freidlin, Dokl. Akad. Nauk. SSSR **3**, 138, 508 (1961).

Dirac constraints in field theory: Lifts of Hamiltonian systems to the cotangent bundle

Walter Oevel
Universitaet Paderborn, D4790 Paderborn, Federal Republic of Germany

(Received 11 June 1987; accepted for publication 26 August 1987)

To every Hamiltonian system is associated a degenerate Lagrangian formulation. Following Dirac's theory of constraints one finds a family of lifts of the system to the cotangent bundle of the underlying manifold. The lifted system admits a reduction to the original equation such that the original Hamiltonian formulation is the pullback of the canonical symplectic form. Invariants of the original equation can be lifted to invariants of the extended system. This procedure is applied to integrable systems such as the Korteweg–de Vries equation.

I. INTRODUCTION

In some recent papers Nutku[1] and Olver[2] considered field equations as Hamiltonian systems (on infinite-dimensional manifolds) for which a completely degenerate Lagrangian formulation is given. This degeneracy arises from the fact that neither the considered spaces are given as cotangent bundles nor is it obvious how to split them into a foliation of Lagrangian submanifolds. So the Dirac machinery[1,2] was invoked to construct an extension of the original system: a new dependent variable—endowed with a suitable time evolution—is introduced. The coupled dynamical system might be called "triangular": the coupling occurs only in the time evolution for the new variable, whereas the dynamic of the original fields remains unchanged. Hence this new field is a gauge field in the sense that it does not interfere with the original variables. By a proper choice of time evolution induced by Dirac's theory the coupled system automatically is Hamiltonian w.r.t. a canonical symplectic structure.

In Ref. 1 this idea was applied to some examples by explicitly using the Dirac machinery. Olver[2] succeeded in simplifying this procedure for the class of dynamical systems, for which the Hamiltonian formulation is given by a field-independent first-order differential operator. The resulting extended system can be expressed explicitly in terms of the original Hamiltonian using the "potential" of the original field variable.

The restriction to field-independent Hamiltonian operators certainly is severe: many Hamiltonian formulations in fluid dynamics[3] arise from Lie–Poisson brackets and hence do depend (linearly) on the fields. Of the infinitely many Hamiltonian formulations of integrable field equations, such as the KdV,[4] usually all up to one depend on the field. Furthermore the geometry of Olver's construction is not obvious, the role of the "potential" variable remains somewhat unclear.

The goal of this paper is to give a geometric interpretation of this procedure. It turns out that the extended system can be given by a straightforward lift of the original dynamical system to the cotangent bundle of the underlying manifold, the new Hamiltonian formulation is induced by the canonical symplectic structure of that cotangent bundle. This lift can be given for any Hamiltonian system, there are no restrictions whatsoever for the Hamiltonian operator.

Although there are certain ambiguities in this procedure (the degenerate Lagrangian system associated to the considered equation is unique only up to a closed one-form) the induced liftings of the Hamiltonian system are more or less the same, they differ only by a simple diffeomorphism: the new "gauge" variable is just shifted by adding this one-form. If different Hamiltonian formulations are known (e.g., in the case of integrable field equations such as the KdV) then the liftings w.r.t. all these Hamiltonian operators again coincide up to a shift in the "gauge" variable. Indeed, the basic part of the lift is given by the standard lift[5] of a vector field to the cotangent bundle, the Hamiltonian formulation of the considered equation only enters into a shift diffeomorphism following this standard lift.

We will investigate the structure of the new coupled system to a certain extent; in particular we will show that integrability of an equation induces integrability of the lifted system. To this end one has to show that invariants (such as conservation laws or recursion operators[6]) can be lifted to invariants of the new system.

The following procedures will apply both to the finite- as well as the infinite-dimensional case. We will first recall the basic framework of differential geometry by giving the definitions in a form that makes sense for the finite-dimensional case as well as for field equations. In Sec. III the lifting of the Hamiltonian system will be given and motivated using Dirac's theory of constraints. In Sec. IV a systematic scheme of lifting all kinds of tensor fields will be proposed, such that the resulting structures are natural w.r.t. composition, exterior derivation, and Lie derivatives. This immediately shows how to lift invariants of the original system. In Sec. V we will discuss the bi-Hamiltonian case of an integrable equation and demonstrate for the KdV example how this structure survives the lifting.

Before finishing this introduction we feel that some critical remarks are due. The proposed lifting of a Hamiltonian system is basically just the standard lift which casts *every* dynamical system into canonical Hamiltonian form. The price one has to pay essentially is the introduction of new variables, so the new system—although in canonical form—is more complicated than the original one. We regard some aspects of the lifting interesting from a geometrical point of view, but we do not know to what extent this procedure might help to simplify the original equation or how to obtain additional information about it. In a certain sense an analog of Darboux's theorem is achieved: having started with a noncanonical formalism one has added canonical variables ending up at the starting point of the classical symplectic machinery, a canonical Hamiltonian system on a cotangent bundle. One of the relevant Hamiltonian operators found by this procedure is the standard symplectic structure, hence the extended system can be regarded as written down in Darboux's coordinates.

On the other hand, ignoring the Lagrangian point of view and sticking to the context of Poisson manifolds, it seems desirable to achieve a Darboux-like "normal form" of the Poisson operators not by adding canonical variables but rather by ignoring irrelevant degrees of freedom (Casimir functions), restricting oneself to the symplectic leaves of the underlying manifold and looking for convenient coordinates there. Thus the system would be simplified *and* a canonical formalism (although not on a cotangent bundle) would be set up, at least locally. In the finite-dimensional case the generalization of Darboux's theorem to such a situation is well known,[7] but such a scheme for infinite-dimensional manifolds is by no means obvious. This paper definitely does not proceed along these lines. We remark that for integrable systems the inverse scattering transform has turned out to provide Darboux coordinates for certain Hamiltonian operators[8] without extending the configuration space; a more sophisticated use of the degenerate Lagrangian description for these cases (see, e.g., Ref. 9) leads to similar results.

II. NOTATION

Let M be a manifold (of finite or infinite dimension), and TM and T^*M be its tangent and cotangent bundle, respectively. For $u \in M$ the tangent and cotangent spaces are denoted by T_uM and T_u^*M. Let $T_p^q(M)$ be the space of p-fold covariant and q-fold contravariant tensor fields. We will fix our notation such that the letters $f \in F(M) \equiv T_0^0(M)$, $K,\tau \in X(M) \equiv T_0^1(M)$, $\gamma,\mu \in X^*(M) \equiv T_1^0(M)$, $J \in T_2^0(M)$, $\Theta \in T_0^2(M)$, $\Phi \in T_1^1(M)$ are reserved for elements of the spaces given above. The elements of $F(M)$ and $X^*(M)$ are the zero- and one-forms on M, $X(M)$ is the space of vector fields. The second-degree tensor fields will be regarded as maps between the tangent and cotangent bundles by identifying J, Θ, and Φ with the tensor fields

$$\begin{aligned}
&T_J(u)(X_1,X_2) = \langle J(u)X_1,X_2\rangle, \quad X_1,X_2 \in T_uM;\\
&T_\Theta(u)(X_1^*,X_2^*) = \langle X_1^*,\Theta(u)X_2^*\rangle, \quad X_1^*,X_2^* \in T_u^*M;\\
&T_\Phi(u)(X,X^*) = \langle X^*,\Phi(u)X\rangle, \quad X \in T_uM, \quad X^* \in T_u^*M.
\end{aligned} \tag{2.1}$$

For example, J will be regarded as a map from the vector fields to the covector fields, Θ as a map from the covector fields to the vector fields, and Φ as a map from the vector fields to the vector fields. The bracket $\langle\cdot,\cdot\rangle$ is to denote the duality between covectors and vectors. All these tensor fields are called nondegenerate if the maps between the fibers of the corresponding bundles are invertible. By the composition of two tensor fields we understand the composition of these maps, e.g., for $J \in T_2^0(M)$, $\Theta \in T_0^2(M)$, we have $\Theta J \in T_1^1(M)$.

As all these objects have to be invariant w.r.t. the choice of a special chart of the manifold, we have to claim the usual transformation laws. Let $\tilde{u} = \tilde{u}(u)$ be a coordinate transformation of M and $\partial\tilde{u}/\partial u$ its Jacobian with inverse $\partial u/\partial\tilde{u}$, then we have, in the chart given by $\tilde{u}$,

$$\begin{aligned}
&\tilde{f}(\tilde{u}) = f(u), \quad f,\tilde{f} \in F(M),\\
&\tilde{K}(\tilde{u}) = \frac{\partial\tilde{u}}{\partial u} K(u), \quad K,\tilde{K} \in X(M),\\
&\tilde{\gamma}(\tilde{u}) = \left(\frac{\partial u}{\partial\tilde{u}}\right)^* \gamma(u), \quad \gamma,\tilde{\gamma} \in X^*(M),\\
&\tilde{J}(\tilde{u}) = \left(\frac{\partial u}{\partial\tilde{u}}\right)^* J(u)\,\frac{\partial u}{\partial\tilde{u}}, \quad J,\tilde{J} \in T_2^0(M),\\
&\tilde{\Theta}(\tilde{u}) = \frac{\partial\tilde{u}}{\partial u}\,\Theta(u)\left(\frac{\partial\tilde{u}}{\partial u}\right)^*, \quad \Theta,\tilde{\Theta} \in T_0^2(M),\\
&\tilde{\Phi}(\tilde{u}) = \frac{\partial\tilde{u}}{\partial u}\,\Phi(u)\frac{\partial u}{\partial\tilde{u}}, \quad \Phi,\tilde{\Phi} \in T_1^1(M).
\end{aligned} \tag{2.2}$$

Here * denotes the transpose w.r.t. the duality $\langle\cdot,\cdot\rangle$. In the infinite-dimensional case the Jacobian will be an operator; nontrivial examples of such coordinate changes are given by the Bäcklund transformations of integrable equations.[10]

The basic tool to handle these various sorts of tensor fields is the notion of Lie derivatives[11] into the direction of a vector field. They may be defined invariantly, but having in mind applications to infinite-dimensional systems it is more convenient to work in local coordinates.

Definition 1: Let E be one of the above spaces. The Lie derivatives $L_\tau: E \to E$ of the following quantities into the direction of $\tau \in X(M)$ are

(a) for $f \in F(M)$, $\quad L_\tau f := \langle df,\tau\rangle := f'[\tau]$;

(b) for $K \in X(M)$, $\quad L_\tau K := [\tau,K] := K'[\tau] - \tau'[K]$;

(c) for $\gamma \in X^*(M)$,

$$L_\tau\gamma := \gamma'[\tau] + \tau'^*\gamma \equiv (\gamma' - \gamma'^*)\tau + d\,\langle\gamma,\tau\rangle;$$

(d) for $J \in T_2^0(M)$, $\quad L_\tau J := J'[\tau] + \tau'^*J + J\tau'$;

(e) for $\Theta \in T_0^2(M)$, $\quad L_\tau\Theta := \Theta'[\tau] - \tau'\Theta - \Theta\tau'^*$;

(f) for $\Phi \in T_1^1(M)$, $\quad L_\tau\Phi := \Phi'[\tau] - \tau'\Phi + \Phi\tau'$.

Here all objects are given by a chart representation, the prime indicates the usual directional derivative of a tensor field $T(u)$, i.e.,

$$T'(u)[X] := \frac{\partial}{\partial\varepsilon} T(u + \varepsilon X)\Big|_{\varepsilon=0}, \quad X \in T_uM. \tag{2.3}$$

One checks that all these definitions are invariant w.r.t. the above transformation laws. These notations coincide with the classical, i.e., finite-dimensional, definitions but they also make sense for a huge class of field equations: in many examples the tensor fields are polynomials of linear operators (differential and integration operators) and expressions of the field variable u and its derivatives or integrations. Hence the directional derivatives imply only differentiations w.r.t. polynomials of ε, which can be done in a naive way without any topological technicalities (which we will ignore, anyway). As long as a chain rule is satisfied for (2.3) and second derivatives are symmetric one finds the usual properties of Lie derivatives, i.e., they are representations of the vector field Lie algebra

$$L_{[\tau,K]} = L_\tau \circ L_K - L_K \circ L_\tau, \tag{2.4}$$

for each of the cases in Definition 1, and one has a general product rule

$$L_\tau (A \circledast B) = (L_\tau A) \circledast B + A \circledast (L_\tau B), \tag{2.5}$$

for any fields A and B in one of the spaces above and meaningful operations $\circledast$ between them. This operation $\circledast$ could be an inner product, the commutator of two vector fields or just the composition of two tensor fields, e.g., $L_\tau(\Theta J) = (L_\tau \Theta) J + \Theta (L_\tau J)$, where one has to choose the proper definition of L_τ according to the differentiated object.

The familiar structure of differential forms shall be introduced into our notation in the following way: The differential $df \in X^*(M)$ of a function $f \in F(M)$ is defined (locally, in a chart) by

$$\langle df(u), X\rangle := f'(u)[X], \quad X \in T_u M. \tag{2.6}$$

For $\gamma \in X^*(M)$ we define its differential $d\gamma \in T^0_2(M)$ as the antisymmetric tensor field, i.e., the two-form

$$d\gamma(u)(X_1,X_2) := \langle \gamma'(u)[X_1], X_2\rangle - \langle \gamma'(u)[X_2], X_1\rangle,$$
$$X_1, X_2 \in T_u M, \tag{2.7}$$

or $d\gamma = \gamma' - \gamma'^*$ for short. For antisymmetric $J \in T^0_2(M)$, i.e., a two-form, we define dJ to be the three-form given by

$$dJ(u)(X_1,X_2,X_3)$$
$$:= \langle J'(u)[X_1]X_2, X_3\rangle + \langle J'(u)[X_2]X_3, X_1\rangle$$
$$+ \langle J'(u)[X_3]X_1, X_2\rangle, \quad X_1, X_2, X_3 \in T_u M. \tag{2.8}$$

According to Poincaré's lemma the closed forms coincide (locally) with the exact forms, e.g., $\gamma \in X^*(M)$ is the gradient of a function f iff it is closed, i.e., $d\gamma = 0$. The same holds for a two-form, i.e., an antisymmetric $J \in T^0_2(M)$ has a potential $\gamma \in X^*(M)$ with $J = d\gamma$ iff $dJ = 0$. In star-shaped regions these potentials can be expressed via contour integrals along straight lines: if $u = 0$ is the star point, then one has

$$d\gamma = 0 \Leftrightarrow \gamma = df \quad \text{with } f(u) = \int_0^1 \langle \gamma(\lambda u), u\rangle d\lambda,$$
$$dJ = 0 \Leftrightarrow J = d\gamma \tag{2.9}$$
$$\text{with } \langle \gamma(u), X\rangle = \int_0^1 \langle J(\lambda u)\lambda u, X\rangle d\lambda, \quad X \in T_u M.$$

Again we note that for most of the infinite-dimensional examples these integrals do not lead to any difficulties, as the objects depend on the field variable and its derivatives in a polynomial way. Of course the potentials are unique only up to an exact form, i.e., for a closed antisymmetric $J \in T^0_2(M)$ we have $J = d(\gamma + df)$, where $f \in F(M)$ is arbitrary and $\gamma \in X^*(M)$ is given by (2.9). It is easily checked that Lie derivatives and exterior derivation commute.

Definition 2: An antisymmetric $\Theta \in T^2_0(M)$ is called a *Poisson operator* if the expression

$$\{\{X^*_1, X^*_2, X^*_3\}\} := \langle X^*_1, \Theta'(u)[\Theta(u)X^*_2]X^*_3\rangle \tag{2.10}$$

satisfies the Jacobi identity for all $X^*_1, X^*_2, X^*_3 \in T^*_u M$. The *Poisson bracket* induced by such an operator is given by

$$\{f,g\}_\Theta := \langle dg, \Theta\, df\rangle, \quad f,g \in F(M). \tag{2.11}$$

One easily checks that the inverse of a symplectic operator is a Poisson operator and vice versa. We will distinguish between two types of Hamiltonian systems: those arising from closed forms and others induced by Poisson operators.

Definition 3: A vector field $K \in X(M)$ is called (a) *Hamiltonian* w.r.t. a closed $J \in T^0_2(M)$, if $JK = df$ for some $f \in F(M)$; or (b) *inverse Hamiltonian* w.r.t. a Poisson operator $\Theta \in T^2_0(M)$, if K is of the form $K = \Theta dg$, $g \in F(M)$.

We do not claim nondegeneracy for the Hamiltonian operators J or Θ, hence both definitions will, in general, not coincide.

A tensor field is called *invariant* w.r.t. $K \in X(M)$ if its Lie derivative into the direction of K vanishes. Obviously invariant zero-forms are conservation laws, the flows of invariant vector fields are one-parameter symmetry groups. An invariant $J \in T^0_2(M)$ maps such symmetry generators to invariant covector fields, an invariant $\Theta \in T^2_0(M)$ works the other way round. For both cases of Definition 3 the Hamiltonian operator, i.e., the closed form or the Poisson operator, respectively, automatically is an invariant of the Hamiltonian system. An invariant $\Phi \in T^1_1(M)$ is called a *recursion operator*[6]; it maps symmetry generators to new symmetry generators.

It is well known[10] that for most of the examples of integrable field equations their symmetry group can be approached by recursion operators. In all these cases the commutativity of the symmetries is reflected by an algebraic property of the recursion operator: its Nijenhuis tensor[5] vanishes. This was called the "hereditary property" in Ref. 10, it can equivalently be given by the following definition.

Definition 4: A tensor field $\Phi \in T^1_1(M)$ is called *hereditary* if $L_{\Phi\tau}\Phi = \Phi L_\tau \Phi$ for all vector fields $\tau \in X(M)$.

If a recursion operator is known for an integrable equation, then this property immediately leads to the construction of commuting symmetries.[12] If this operator is given as the composition of two Hamiltonian structures, then this property reflects a compatibility between the Hamiltonian operators leading to the construction of a hierarchy of conservation laws in involution.[10,13,14]

III. DIRAC CONSTRAINTS AND STANDARD LIFTS

We first want to generalize the procedure of Ref. 2: Consider a vector field $K \in X(M)$ that is Hamiltonian w.r.t. a closed (and maybe degenerate) $J \in T^0_2(M)$, i.e.,

$$JK = df, \quad f \in F(M). \tag{3.1}$$

As J is closed it locally can be written as $J = -d\mu_0 = \mu_0'^* - \mu_0'$, where μ_0 can be obtained from (2.9). The dynamical system $u_t = K(u)$ obviously can be regarded as the Euler–Lagrange equation

$$\frac{d}{dt}\frac{\partial L}{\partial u_t} = \frac{\partial L}{\partial u} \tag{3.2}$$

for the Lagrangian function

$$L(u,u_t) := \langle \mu_0(u), u_t \rangle - f(u). \tag{3.3}$$

As usual a Lagrangian for a given system is determined only up to a closed one-form, this ambiguity here is reflected by the fact that the potential μ_0 of the Hamiltonian operator J is given only up to a closed one-form. Of course, as we have not taken into account any decomposition of the underlying manifold this artificially constructed Lagrangian system is completely degenerate. Following Refs. 1 and 2 one formally introduces the variable π as conjugate of u by

$$\pi := \frac{\partial L}{\partial u_t}, \tag{3.4}$$

such that one encounters a Dirac constraint $\pi = \mu_0(u)$. One now tries to obtain the original Hamiltonian system as a reduction of a canonical Hamiltonian system in the variables u and π (being coordinates of the cotangent bundle) by this constraint. The total Hamiltonian H of this extended system is obtained from the original f by adding the constraints via Lagrangian multipliers Λ, i.e.,

$$H(u,\pi) = f(u) + \langle \pi - \mu_0(u), \Lambda(u,\pi) \rangle. \tag{3.5}$$

The multipliers are to be determined by the fact that the extended Hamiltonian system (w.r.t. to the canonical symplectic form on T^*M) can be reduced to the original equation, i.e., on the submanifold $\pi = \mu_0(u)$ the Hamiltonian has to be involution with the constraints. Hence

$$\{H, \langle \pi - \mu_0, \tau \rangle\}_{\text{can}} = 0 \tag{3.6}$$

has to be satisfied for an arbitrary vector field $\tau \in X(T^*M)$. This Poisson bracket is induced by the Poisson operator Θ_{can} given by the inverse of the canonical symplectic form J_{can} on T^*M, i.e.,

$$\begin{aligned} J_{\text{can}}(u,\pi) &= \begin{pmatrix} 0 & ; & -1^* \\ 1 & ; & 0 \end{pmatrix}, \\ \Theta_{\text{can}}(u,\pi) &= \begin{pmatrix} 0 & ; & 1 \\ -1^* & ; & 0 \end{pmatrix}, \\ \{H_1,H_2\}_{\text{can}} &= \langle dH_2, \Theta_{\text{can}}\, dH_1 \rangle \\ &= \left\langle \frac{\partial H_2}{\partial u}, \frac{\partial H_1}{\partial \pi} \right\rangle - \left\langle \frac{\partial H_1}{\partial u}, \frac{\partial H_2}{\partial \pi} \right\rangle, \end{aligned} \tag{3.7}$$

with $H_1, H_2 \in F(T^*M)$, and 1 and 1^* being the identity maps on the fibers of TM and T^*M. Inserting (3.5) into (3.6) one sees that

$$\langle df + \mu_0' \Lambda - \mu_0'^* \Lambda, \tau \rangle \tag{3.8}$$

has to vanish for any $\tau \in X(T^*M)$. As $df = JK = (\mu_0'^* - \mu_0')K$ one finds that with the choice $\Lambda = K$ the extended system admits the required reduction. For degenerate J elements of its kernel might be added to Λ. Choosing $\Lambda = K$ one calculates the Hamiltonian system associated to (3.5) and finds a natural lifting of the vector field K to the cotangent bundle of M.

Proposition 1: Let (u,π) be local coordinates of T^*M. To every dynamical system $u_t = K(u)$ on M we associate the lifted system

$$\begin{aligned} \frac{d}{dt}\begin{pmatrix} u \\ \pi \end{pmatrix} = l_0(K;\mu_0) &:= \begin{pmatrix} K \\ -K'^*\pi + L_K\mu_0 \end{pmatrix} \\ &= \Theta_{\text{can}}\left(d\langle \pi, K \rangle - \begin{pmatrix} L_K\mu_0 \\ 0 \end{pmatrix} \right) \end{aligned} \tag{3.9}$$

on T^*M. This system admits the reduction $\pi = \mu_0(u)$, i.e., the lifted vector field is tangent to the graph of μ_0 in T^*M. If K is Hamiltonian w.r.t. $J = -d\mu_0$, i.e., $JK = df$ for some $f \in F(M)$, then

$$L_K\mu_0 \equiv d\langle \mu_0, K \rangle - JK = d(\langle \mu_0, K \rangle - f).$$

Hence the lifting of a Hamiltonian vector field is Hamiltonian w.r.t. the canonical symplectic form on T^*M with Hamiltonian function $f + \langle \pi - \mu_0, K \rangle$.

The reduction property of the lift is calculated easily. Although motivated by Dirac's theory, it does not depend on the Hamiltonian form of K.

Remark 1: The closed operator $J = -d\mu_0$ is the pullback of the canonical form J_{can} via the map $u \in M \rightarrow (u, \mu_0(u)) \in T^*M$.

Remark 2: The lift $l_0(K;\mu_0)$ is *always* Hamiltonian w.r.t. to the symplectic operator

$$J_{\mu_0}(u,\pi) := \begin{pmatrix} d\mu_0 & ; & -1^* \\ 1 & ; & 0 \end{pmatrix} \tag{3.10a}$$

and its inverse

$$\Theta_{\mu_0}(u,\pi) := \begin{pmatrix} 0 & ; & 1 \\ -1^* & ; & d\mu_0 \end{pmatrix}. \tag{3.10b}$$

One finds

$$J_{\mu_0} l_0(K;\mu_0) = d\langle \pi - \mu_0, K \rangle. \tag{3.11}$$

So the above lift turns every dynamical system into a Hamiltonian one; Hamiltonian systems are cast into bi-Hamiltonian form. We remark that this bi-Hamiltonian formulation is very trivial compared to the structures found for integrable systems,[10] as the corresponding hereditary recursion operator

$$\begin{aligned} \Phi_{\mu_0} &:= \Theta_{\mu_0} J_{\text{can}} = \begin{pmatrix} 1 & ; & 0 \\ d\mu_0 & ; & 1^* \end{pmatrix}, \\ \Theta_{\text{can}} J_{\mu_0} &= \begin{pmatrix} 1 & ; & 0 \\ -d\mu_0 & ; & 1^* \end{pmatrix} = \Phi_{\mu_0}^{-1} \end{aligned} \tag{3.12}$$

is just the identity map plus a nilpotent operator.

The lift $l_0(K;\mu_0)$ reminds us of the standard (or complete[5]) lift

$$\begin{pmatrix} K \\ -K'^*\pi \end{pmatrix} = \Theta_{\text{can}} d\langle \pi, K \rangle \tag{3.13}$$

of a vector field K on M to T^*M, which casts every vector field into canonical form. Indeed, if one considers the shift diffeomorphism on T^*M given in local coordinates by

$$\begin{pmatrix} u \\ \pi \end{pmatrix} \rightarrow \begin{pmatrix} u \\ \pi + \mu_0(u) \end{pmatrix}, \tag{3.14}$$

then it is easily calculated that $l_0(K;\mu_0)$ is obtained as the pushforward of (3.13) via (3.14). There also is a standard lift of tensor fields $\Phi \in T_1^1(M)$ to T^*M (the "complete" lift in Ref. 5). Composing this lift with the transformation (3.14), i.e., using (2.2) on T^*M, one obtains a lift

$$l_0(\Phi;\mu_0):=\begin{pmatrix}\Phi(u) & ; & 0\\ \hat{\Phi}(u,\pi) & ; & \Phi^*(u)\end{pmatrix}, \tag{3.15}$$

where $\hat{\Phi}(u,\pi)$ is given by

$$\hat{\Phi}=\Phi^{*\prime}[\cdot]\pi-(\Phi^*[\cdot]\pi)^*-d(\Phi^*\mu_0)+(d\mu_0)\Phi,$$

i.e.,

$$\langle\hat{\Phi}X_1,X_2\rangle=\langle\pi-\mu_0,\Phi'[X_1]X_2-\Phi'[X_2]X_1\rangle$$
$$+\langle\mu_0'[\Phi X_1],X_2\rangle-\langle\mu_0'[X_1],\Phi X_2\rangle,$$

with $X_1,X_2\in T_uM$.

The definitions of the lift l_0 are invariant w.r.t. cotangent bundle transformations

$$\tilde{u}=\tilde{u}(u),\quad \tilde{\pi}=\left(\frac{\partial u}{\partial\tilde{u}}\right)^*\pi, \tag{3.16}$$

for invariant definitions of the "complete" part of these lifts see Ref. 5. The additional shift $\pi\rightarrow\pi+\mu_0$ certainly does not affect the invariance.

These lifts are natural w.r.t. Lie derivatives, one finds

$$\begin{gathered}[l_0(K;\mu_0),l_0(\tilde{K};\mu_0)]=l_0([K,\tilde{K}];\mu_0),\\ L_{l_0(K;\mu_0)}l_0(\Phi;\mu_0)=l_0(L_K\Phi;\mu_0),\end{gathered} \tag{3.17}$$

for arbitrary $K,\tilde{K}\in X(M)$ and $\Phi\in T_1^1(M)$. After some lengthy calculations one finds that the Nijenhuis tensor[5] of $l_0(\Phi;\mu_0)$ is related to the Nijenhuis tensor of Φ: a hereditary Φ on M is lifted to a hereditary operator on T^*M. The lift l_0 is not natural w.r.t. composition of tensor fields, one finds

$$l_0(\Phi;\mu_0)l_0(K;\mu_0)=l_0(\Phi K;\mu_0)$$
$$+\begin{pmatrix}0\\ (L_K\Phi)^*(\pi-\mu_0)\end{pmatrix}, \tag{3.18}$$

so the liftings commute with composition only with additional invariance assumptions. Similarly, only for hereditary Φ does a power of the lifted operator $l_0(\Phi;\mu_0)$ coincide with the lift of the power of Φ (see Ref. 5).

As Lie derivatives commute with l_0, one immediately can lift all the known generators of symmetries for $K\in X(M)$ to T^*M. A (hereditary) recursion operator can be lifted to a (hereditary) recursion operator for $l_0(K;\mu_0)$. By a straightforward calculation following compatibility structures between the lifted operator and the natural symplectic forms J_{can} and J_{μ_0} on T^*M are found.

Remark 3: For every $\Phi\in T_1^1(M)$ the operator

$$J_{\mu_0}l_0(\Phi;\mu_0)=d\begin{pmatrix}\Phi^*(\pi-\mu_0)\\ 0\end{pmatrix}$$

is exact and hence closed. If $(d\mu_0)\Phi$ is closed, then $J_{\text{can}}l_0(\Phi;\mu_0)$ is also closed.

A compatibility with the natural Poisson structures Θ_{can} and Θ_{μ_0} on T^*M is granted for hereditary operators.

Remark 4: For every hereditary $\Phi\in T_1^1(M)$ the operator $l_0(\Phi;\mu_0)\Theta_{\mu_0}$ is Poisson. This implies that this operator is compatible with Θ_{μ_0} in the sense of Refs. 10, 13, and 14, i.e., the sum of these operators is again Poisson. If, in addition, $(d\mu_0)\Phi$ is closed, then it is checked that $l_0(\Phi;\mu_0)\Theta_{\text{can}}$ as well as $\Theta_{\text{can}}+l_0(\Phi;\mu_0)\Theta_{\text{can}}$, $\Theta_{\mu_0}+l_0(\Phi;\mu_0)\Theta_{\text{can}}$, and $\Theta_{\text{can}}+l_0(\Phi;\mu_0)\Theta_{\mu_0}$ are also Poisson.

The lifted operators turn out to be symplectically self-adjoint w.r.t. J_{μ_0}, i.e., for any $\Phi\in T_1^1(M)$ one finds

$$l_0(\Phi;\mu_0)^*=J_{\mu_0}l_0(\Phi;\mu_0)\Theta_{\mu_0}. \tag{3.19}$$

If $(d\mu_0)\Phi=\Phi^*(d\mu_0)$, then one also has

$$l_0(\Phi;\mu_0)^*=J_{\text{can}}l_0(\Phi;\mu_0)\Theta_{\text{can}}, \tag{3.20}$$

implying that $l_0(\Phi;\mu_0)$ and Φ_{μ_0} commute.

Before further investigation of the lift (3.9) we will consider the following example.

The KdV example: As manifold we choose the Schwartz space $M=S(\mathbb{R})$ of functions on the real line vanishing rapidly at infinity. We identify the tangent spaces T_uM with $S(\mathbb{R})$ and embed them into their duals via the pairing

$$\langle a^*,a\rangle=\int_{-\infty}^{\infty}a^*(x)a(x)dx. \tag{3.21}$$

The differential operator is denoted by D, its "inverse" is given by

$$(D^{-1}a)(x)=\int_{-\infty}^{x}a(\xi)d\xi-\frac{1}{2}\int_{-\infty}^{\infty}a(\xi)d\xi. \tag{3.22}$$

With $a\in S(\mathbb{R})$ and $a^*\in D^{-1}(S(\mathbb{R}))\subset S^*(\mathbb{R})$ one checks $DD^{-1}a=a$ and $D^{-1}Da^*=a^*$. We will restrict the cotangent spaces to the subset $D^{-1}(S(\mathbb{R}))$, so the differential operator, interpreted as element of $T_0^2(M)$, is indeed invertible. Due to the boundary conditions both D and D^{-1} are antisymmetric; hence according to the definitions (2.10) and (2.8) D is a Poisson operator and D^{-1} is closed. These operators induce a Hamiltonian formulation for a large class of field equations, among them the famous Korteweg–de Vries equation

$$u_t=u_{xxx}+6uu_x=Dd\int_{-\infty}^{\infty}\left(u^3-\frac{1}{2}u_x^2\right)dx. \tag{3.23}$$

This inverse Hamiltonian formulation with the Poisson operator D is converted to a Hamiltonian formulation w.r.t. the closed operator $J(u)=D^{-1}$ by integrating (3.23). Using (2.9) one finds a potential $\mu_0(u)=-\frac{1}{2}D^{-1}u$ for J (i.e., $J=-d\mu_0$) and the lift (3.9) yields

$$\begin{pmatrix}u_{xxx}+6uu_x\\ \pi_{xxx}+6u\pi_x+\frac{3}{2}u^2\end{pmatrix}=\begin{pmatrix}0 & ; & 1\\ -1^* & ; & 0\end{pmatrix}d\int_{-\infty}^{\infty}\left(\pi(u_{xxx}+6uu_x)-\frac{1}{2}u^3\right)dx, \tag{3.24}$$

which admits the reduction $\pi=-\frac{1}{2}D^{-1}u$ to the KdV equation (3.23). Using (3.15), the well known hereditary recursion operator[6,10]

$$\Phi(u)=D^2+4u+2u_xD^{-1} \tag{3.25}$$

of the KdV equation is lifted to a hereditary recursion operator

$$l_0(\Phi;\mu_0)=\begin{pmatrix} D^2+4u+2u_xD^{-1} & ; & 0 \\ D^{-1}(2\pi_x+u)+(2\pi_x+u)D^{-1} & ; & D^2+4u-2D^{-1}u_x \end{pmatrix} \tag{3.26}$$

for (3.24). According to remarks 3 and 4 the composition of this operator with the Poisson operators Θ_{can} and Θ_{μ_0} will yield new Poisson operators, which are natural candidates for nontrivial multi-Hamiltonian formulations of (3.24). This will be discussed in Sec. V.

In Ref. 2 the occurrence of the integration operator J was avoided by mapping the original equation to its "potential form" first: introducing $\phi_x=u$ and using the transformation laws (2.2) one finds that the "potential form" admits the closed operator $\tilde{J}(\phi)=(-D)D^{-1}D=-D$ as Hamiltonian structure. Indeed, (3.23) is transformed to

$$\begin{aligned} &\phi_t=\phi_{xxx}+3\phi_x^2, \\ &(-D)(\phi_{xxx}+3\phi_x^2)=d\int_{-\infty}^{\infty}\left(\phi_x^3-\frac{1}{2}\phi_{xx}^2\right)dx. \end{aligned} \tag{3.27}$$

Using the potential $\tilde{\mu}_0(\phi)=\frac{1}{2}\phi_x$ (i.e., $\tilde{J}=-d\tilde{\mu}_0$) one obtains the lift

$$\begin{aligned} &\begin{pmatrix} \phi_{xxx}+3\phi_x^2 \\ \pi_{xxx}+6\phi_x\pi_x+6\phi_{xx}\pi-3\phi_x\phi_{xx} \end{pmatrix} \\ &\quad=\begin{pmatrix} 0 & ; & 1 \\ -1^* & ; & 0 \end{pmatrix} d\int_{-\infty}^{\infty}\Big(\pi(\phi_{xxx}+3\phi_x^2) \\ &\quad-\frac{1}{2}\phi_x^3\Big)dx, \end{aligned} \tag{3.28}$$

admitting the reduction $\pi=\frac{1}{2}\phi_x$ to the potential KdV equation (3.27). Up to an irrelevant rescaling this coincides with the results of Ref. 2. In general, it is easily demonstrated that Theorem 5 of Ref. 2 is the special case of Proposition 1 with $\mu_0(u)=-\frac{1}{2}D^{-1}u$.

Using (2.2) one transforms (3.25) into a hereditary recursion operator

$$\Phi(\phi)=D^2+4\phi_x-2D^{-1}\phi_{xx} \tag{3.29}$$

for the potential KdV equation (3.27), which is lifted to a hereditary recursion operator

$$\begin{pmatrix} D^2+4\phi_x-2D^{-1}\phi_{xx} & ; & 0 \\ D(2\pi-\phi_x)+(2\pi-\phi_x)D & ; & D^2+4\phi_x+2\phi_{xx}D^{-1} \end{pmatrix}$$

for (3.28).

IV. LIFTINGS FOR VARIOUS TENSOR FIELDS

We now want to investigate the structure of the lifted dynamical system (3.9). Much information about an equation is obtained by its invariants, i.e., tensor fields with vanishing Lie derivatives into the direction of the considered vector field. We want to find liftings of invariants on M to T^*M, such that the lifted structures are invariants for (3.9). For invariant vector fields (i.e., symmetry generators) and invariant tensor fields of type (1,1) (i.e., recursion operators) this is achieved by the lifting l_0. For other kinds of tensor fields there does not seem to be an analogy for this direct lift. Nevertheless, for Hamiltonian systems there is a way of obtaining such liftings, as the Hamiltonian operator (i.e., either a closed form or a Poisson structure) yields a transformation between tangent and cotangent bundles. Hence it is convenient to obtain the desired lifts to T^*M in three steps.

(1) There are natural liftings of all kinds of tensor fields from M to its tangent bundle TM.

(2) If a map between TM and T^*M is given (e.g., induced by a Hamiltonian operator) then the lifts on TM can be pushed forward or pulled back to the cotangent bundle.

(3) To obtain the lifting (3.9) motivated by Dirac's theory, one just has to apply the shift diffeomorphism (3.14) on the fibers of T^*M to the results of the first two steps.

We remark that step (3) is not at all important for the geometric concept behind this construction, it is just motivated by the considerations leading to (3.9).

The lifts to the tangent bundle are the "complete" lifts of Ref. 5. They may be defined invariantly, but aiming for infinite-dimensional examples it is more convenient to give them in a local coordinate frame. These definitions are given in the Appendix.

This lifting l_1, say, is natural w.r.t. composition, Lie derivation, and exterior derivation, i.e.,

$$\begin{aligned} &l_1(T)\circledast l_1(\tilde{T})=l_1(T\circledast\tilde{T}), \\ &L_{l_1(K)}l_1(T)=l_1(L_KT),\quad dl_1(\alpha)=l_1(d\alpha). \end{aligned} \tag{4.1}$$

Here K is an arbitrary vector field on M, T and $\tilde{T}$ are any tensor fields on M with $\circledast$ being the composition or inner product, and α is an arbitrary form on M.

For the second step we now fix two tensor fields $\Theta_0\in T_0^2(M)$ and $J_0\in T_2^0(M)$. In the applications these structures will be given by Hamiltonian operators, but the following considerations will hold for arbitrary tensor fields of proper type. If (u,v) and (u,π) are local coordinates of TM and T^*M, respectively, then Θ_0 and J_0 induce maps

$$\begin{aligned} &\bar{\Theta}_0\colon\ T^*M\to TM\colon \begin{pmatrix} u \\ \pi \end{pmatrix}\to\begin{pmatrix} u \\ \Theta_0(u)\pi \end{pmatrix}, \\ &\bar{J}_0\colon\ TM\to T^*M\colon \begin{pmatrix} u \\ v \end{pmatrix}\to\begin{pmatrix} u \\ J_0(u)v \end{pmatrix}, \end{aligned} \tag{4.2}$$

which will be used to transport the lifts l_1 on TM to T^*M. Covariant tensor fields such as $l_1(f)\in F(TM)$, $l_1(\gamma)\in X^*(TM)$, or $l_1(J)\in T_2^0(TM)$ are pulled back via Θ_0 to covariant tensor fields on T^*M. Contravariant objects such as $l_1(K)\in X(TM)$ or $l_1(\Theta)\in T_0^2(TM)$ are pushed forward by $\bar{J}_0$ to $\text{Im}(\bar{J}_0)\subset T^*M$. For simplicity we will assume that J_0 is not degenerate, i.e., $\bar{J}_0$ is to be a diffeomorphism from TM to T^*M, such that contravariant tensor fields can be pushed forward to the whole of T^*M. Now also mixed tensor fields such as $l_1(\Phi)\in T_1^1(TM)$ can be mapped to fields on T^*M.

We now compose the lift l_1 from M to TM with the maps (4.2) connecting TM and T^*M and obtain a lifting from M to T^*M for all the types of tensor fields discussed before.

For step (3) we then fix $\mu_0 \in X^*(M)$ and consider the shift (3.14), which gives rise to transformations of tensor fields on T^*M. The composition of the lifts obtained from the first two steps with these transformations is given by the lift l_2, say, which is parametrized by μ_0 and Θ_0 or J_0, respectively. Using (2.2) it is a straightforward calculation to obtain the proper form of l_2 for all kinds of tensor fields, the results are given in the Appendix.

Due to the construction these liftings are natural. Hence for differential forms, e.g., $f \in F(M)$, $\gamma \in X^*(M)$, or antisymmetric $J \in T_2^0(M)$ one immediately concludes that exterior derivation commutes with the lifting, e.g., $l_2(J;\Theta_0,\mu_0)$ is a closed operator on T^*M if J is closed. Similarly, a Poisson operator $\Theta \in T_0^2(M)$ is lifted to a Poisson operator, a hereditary operator $\Phi \in T_1^1(M)$ yields a hereditary $l_2(\Phi;J_0,\mu_0)$.

For $\Theta_0 = J_0^{-1}$ the construction is also natural w.r.t. the duality between covariant and contravariant tensor fields, i.e., composition of tensor fields commutes with the lifting l_2. In this case obviously Lie derivatives also commute with l_2 for all types of tensor fields considered here.

From the explicit form of l_2 one sees that $l_2(K;J_0,\mu_0)$ coincides with the lift $l_0(K;\mu_0)$ given by (3.9), if J_0 is invariant w.r.t. the vector field K. If J_0 is given by a symplectic form, then for all Hamiltonian vector fields w.r.t. J_0 these two liftings coincide. Hence for Hamiltonian vector fields K w.r.t. a symplectic J_0 it is convenient to regard (3.9) as a special case of $l_2(K;J_0,\mu_0)$. With this interpretation one can immediately lift all invariants of K to obtain invariants for $l_0(K;\mu_0) = l_2(K;J_0,\mu_0)$. Of course, for invariant vector fields and recursion operators now both lifts will yield invariants, as both lifts l_0 and l_2 commute with Lie derivatives.

It turns out that in certain situations the lifts l_2 are related to the lifts l_0. For vector fields the invariance of J_0 leads to coinciding lifts. For (1,1)-tensor fields one finds the following.

Remark 5: Let $J_0 \in T_2^0(M)$ be symplectic. Assume $\Phi^* J_0 = J_0 \Phi$ for an operator $\Phi \in T_1^1(M)$. If $J_0\Phi$ is closed, then $l_2(\Phi;J_0,\mu_0) = l_0(\Phi;\mu_0)$.

For closed forms one has the following remark.

Remark 6: If $\gamma \in X^*(M)$ and $J \in T_2^0(M)$ are closed, then $l_2(\gamma;\Theta_0,\mu_0) = J_{\mu_0} l_0(\Theta_0^*\gamma;\mu_0)$ and $l_2(J;\Theta_0,\mu_0) = J_{\mu_0} l_0 \times (\Theta_0^* J;\mu_0)$. Here $\Theta_0 \in T_0^2(M)$ is arbitrary, J_{μ_0} is given by (3.10a).

For doubly contravariant tensor fields one finds the following.

Remark 7: Let $J_0 \in T_2^0(M)$ be symplectic. If for antisymmetric $\Theta \in T_0^2(M)$ the operator $J_0 \Theta J_0 \in T_2^0(M)$ is closed, then $l_2(\Theta;J_0,\mu_0) = -l_0(\Theta J_0;\mu_0)\Theta_{\mu_0}$. Here Θ_{μ_0} is given by (3.10b).

All these statements are verified by a straightforward computation. An important aspect of remark 7 is that, in this situation, the lift $l_2(\Theta;J_0,\mu_0)$ does not depend on the invertibility of J_0, i.e., the lift l_2 to the image of J_0 admits a natural continuation to the whole of T^*M. Remarks 6 and 7 immediately lead to the following special case.

Remark 8: If $J \in T_2^0(M)$ is symplectic, then $l_2(J;J^{-1},\mu_0) = -J_{\mu_0}$. For an invertible Poisson operator $\Theta \in T_0^2(M)$ one finds $l_2(\Theta;\Theta^{-1},\mu_0) = -\Theta_{\mu_0}$.

Another remarkable relation between the lifts l_0 and l_2 is given by the Lie derivative of a tensor field obtained by l_2 into the direction of a vector field obtained by l_0.

Remark 9: For a differential form α on M, i.e., either $\alpha = f \in F(M)$ or $\alpha = \gamma \in X^*(M)$ or $\alpha = J \in T_2^0(M)$ (with J being antisymmetric), one finds

$$L_{l_0(K;\mu_0)} l_2(\alpha;\Theta_0,\mu_0) = l_2(L_K\alpha;\Theta_0,\mu_0) + l_2(\alpha;L_K\Theta_0,\mu_0).$$

Corresponding (although more complicated) relations for Lie derivatives of lifts of contravariant or mixed tensor fields can be established.

V. BI-HAMILTONIAN SYSTEMS

We now want to discuss the lifting of bi-Hamiltonian systems.[10,13,14] Such a structure essentially consists of a Poisson operator $\Theta_1 \in T_0^2(M)$ and a closed (but not necessarily symplectic) form $J_0 \in T_2^0(M)$, for which a compatibility condition is given. If J_0 is symplectic, this compatibility is the condition that the sum of Θ_1 and J_0^{-1} is again Poisson.[14] For noninvertible J_0 the same structure is obtained either by claiming the hereditary property[10] for the operator $\Phi = \Theta_1 J_0$ or assuming $J_0\Theta_1 J_0$ to be closed. The best known example of such a Hamiltonian pair[4] is given by

$$\Theta_1(u) = D^3 + 2Du + 2uD, \quad J_0(u) = D^{-1}, \tag{5.1}$$

yielding the hereditary recursion operator (3.25) of the KdV equation. It is well known that (3.23) admits two Hamiltonian formulations

$$\begin{aligned} &K_1 = \Theta_1\, df_0, \quad J_0 K_1 = df_1; \quad K_1(u) := u_{xxx} + 6uu_x, \\ &f_0 := \int_{-\infty}^{\infty} \frac{1}{2} u^2\, dx, \quad f_1 = \int_{-\infty}^{\infty} \left(u^3 - \frac{1}{2} u_x^2\right) dx. \end{aligned} \tag{5.2}$$

This immediately leads to the construction of an integrable system.

Theorem 1 (Refs. 4, 10, 13, and 14): Let $J_0 \in T_2^0(M)$ be closed and $\Theta_1 \in T_0^2(M)$ be Poisson. If $J_0\Theta_1 J_0$ is closed, then $\Phi := J_0\Theta_1$ is hereditary, all operators $J_l := J_0\Phi^l$ are closed and all operators $\Theta_{l+1} := \Phi^l\Theta_1$ are Poisson. If a vector field $K_1 \in X(M)$ is bi-Hamiltonian w.r.t. J_0 and Θ_1, i.e., $K_1 = \Theta_1\, df_0$ and $J_0 K_1 = df_1$ for some functions $f_0, f_1 \in F(M)$, then all the forms $J_l K_1$ are closed. Hence there exist (locally) functions f_l, such that $J_l K_k = df_{k+l}$ and $K_{l+k} = \Theta_k\, df_l$ with $K_{l+1} := \Phi^l K_1$, $k = 1,2,\ldots$, $l = 0,1,\ldots$. All these functions are in involution w.r.t. all the Poisson brackets induced by the Poisson operators Θ_k.

The lifting procedures discussed before can be applied to such a structure in a very simple way: We choose μ_0 such that $J_0 = -d\mu_0$ (or $J_{k_0} = -d\mu_0$ for some $k_0 \in \mathrm{N}$) and define

$$\begin{aligned} &\widetilde{\Phi} := l_0(\Phi;\mu_0), \quad \widetilde{K}_k := \widetilde{\Phi}^k \widetilde{K}_k = l_0(K_k;\mu_0), \\ &\tilde{f}_k := -\langle \pi - \mu_0, K_k \rangle, \quad k = 1,2,\ldots; \\ &\widetilde{\Theta}_l := -\widetilde{\Phi}^l \Theta_{\mu_0}, \quad \widehat{\Theta}_l := \widetilde{\Phi}^l \Theta_{\mathrm{can}}, \\ &\tilde{J}_l := -J_{\mu_0}\widetilde{\Phi}^l, \quad \hat{J}_l := J_{\mathrm{can}}\widetilde{\Phi}^l, \quad l = 0,1,\ldots. \end{aligned} \tag{5.3}$$

As all vector fields K_k are Hamiltonian w.r.t. all the closed operators J_k the liftings $\widetilde{K}_k$ can be regarded as the lift l_2 of Sec. IV. According to remark 5 all the lifts $l_2(\Phi;J_l,\mu_0)$ yield $\widetilde{\Phi}$. The operators $\widetilde{J}_l$ are closed according to remark 3, they can be regarded as the lifts of the closed forms J_l via the Poisson operators Θ_k: according to remark 6 one finds

$$l_2(J_l;\Theta_k,\mu_0) = -J_{\mu_0} l_0(\Theta_k^* J_l;\mu_0) = \widetilde{J}_{k+l}.$$

Similarly the operators $\widetilde{\Theta}_l$ are all Poisson; according to remark 7 one finds $\widetilde{\Theta}_{k+l} = l_2(\Theta_k;J_l,\mu_0)$. The functions $\tilde{f}_k$ obviously are the lifts of the original conservation laws f_l via the Poisson operators Θ_k:

$$\begin{aligned} l_2(f_l;\Theta_k,\mu_0) &= \langle df_l,\Theta_k(\pi-\mu_0)\rangle \\ &= -\langle \pi-\mu_0,\Theta_k df_l\rangle = \tilde{f}_{k+l}. \end{aligned}$$

Due to the natural construction of the lift l_2 the extended functions $\tilde{f}_k$ as well as the operators $\widetilde{\Phi}$, $\widetilde{\Theta}_l$, and $\tilde{J}_l$ are invariants for all the lifted systems $\widetilde{K}_k$ and hence give rise to many different Hamiltonian formulations. Due to the compatibility structure of the operators involved in a bi-Hamiltonian system all lifts l_2 can be expressed in terms of the lifting l_0, hence the lifts l_2 of the Poisson operators to the image of the closed J_l's can be extended to the whole of T^*M and no invertibility assumptions for the operators J_l are needed. Indeed, according to Proposition 1, (3.17), remarks 2–4, one directly obtains the desired invariance of the operators $\widetilde{\Phi}, \widetilde{J}_l$, $\widetilde{J}_l$, $\widetilde{\Theta}_l$, and $\widehat{\Theta}_l$ as well as the fact that they are hereditary, closed, and Poisson, respectively.

Using Proposition 1 and $d\tilde{f}_k = -J_{\mu_0}\widetilde{K}_k$ (remark 2) one easily establishes the Hamiltonian formulations of the lifted dynamical systems:

$$\begin{aligned} &\widetilde{J}_l\widetilde{K}_k = d\tilde{f}_{k+l}, \\ &\hat{J}_l\widetilde{K}_k = d(f_{k+l}-\tilde{f}_{k+l}), \quad (\tilde{J}_l+\hat{J}_l)\widetilde{K}_k = df_{k+l}; \\ &\widetilde{K}_{k+l} = \widetilde{\Theta}_k\, d\tilde{f}_l, \\ &\widetilde{K}_{k+l} = \widehat{\Theta}_k\, d(f_l-\tilde{f}_l), \quad k=1,2,..., \quad l=0,1,... . \end{aligned} \tag{5.4}$$

Because of the "triangular" form of the lifted systems the original Hamiltonian functions f_k are still conservation laws. All the functions are in involution w.r.t. all the Poisson brackets induced by the above Poisson operators. For example, to show that the functions f_l and $\tilde{f}_k$ are in involution w.r.t. the bracket induced by $\widehat{\Theta}_m$, say, one calculates

$$\begin{aligned} \langle df_l,\widehat{\Theta}_m d\tilde{f}_k\rangle &= \langle (\widetilde{J}_0+\hat{J}_0)K_l,\widehat{\Theta}_m\, d\tilde{f}_k\rangle \\ &= \langle d\tilde{f}_1,\widetilde{\Theta}_{l-1}(\widetilde{J}_0+\hat{J}_0)\widehat{\Theta}_m(\widetilde{\Phi}^{k-1})^*\, d\tilde{f}_1\rangle. \end{aligned} \tag{5.5}$$

But according to (3.20) the operator

$$\widetilde{\Theta}_{l-1}(\widetilde{J}_0+\hat{J}_0)\widehat{\Theta}_m(\widetilde{\Phi}^{k-1})^* = \widetilde{\Phi}^{k+l+m-2}(\widetilde{\Theta}_0+\widehat{\Theta}_0)$$

is antisymmetric. Hence the above bracket equals zero. With similar arguments one shows the vanishing of all brackets.

The recursion operators induced by these Hamiltonian formulations are essentially $\widetilde{\Phi}$, $\Phi_{\mu_0} = -\widetilde{\Theta}_0\hat{J}_0$, and its inverse $\widehat{\Theta}_0\widetilde{J}_0$. According to (3.20) these recursion operators commute; remark 4 grants a compatibility between them. Obviously $\widetilde{\Phi}+\Phi_{\mu_0}$ and $\widetilde{\Phi}+\Phi_{\mu_0}^{-1}$ are again hereditary. It can be checked that this together with their commutativity leads to the conclusion that arbitrary polynomials in $\widetilde{\Phi}$, Φ_{μ_0}, and $\Phi_{\mu_0}^{-1}$ are also hereditary. Hence further commuting symmetries can be constructed by applying Φ_{μ_0} to the lifted vector fields $\widetilde{K}_k$.

VI. CONCLUSIONS

We have generalized the procedure of Refs. 1 and 2, which now can be applied in a straightforward way to any Hamiltonian system without restrictions to the explicit form of the Hamiltonian operator. Starting from the Lagrangian point of view, i.e., following Dirac's theory of constrained system, these considerations are motivated in Ref. 2 as a construction to cast a Hamiltonian formulation into canonical form. Further variables are added; the new configuration space is the cotangent bundle of the original manifold, on which the dynamical system assumes canonical Hamiltonian form.

Although it is not clear to us how the lifted systems proposed here might be used to obtain additional information about the original equations, we nevertheless regard the lifting procedure as interesting from a geometrical point of view. The most remarkable observation seems to be that the lifting depends on the Hamiltonian structure only in a very simple way: whatever Hamiltonian operator is chosen for the lift, the resulting equations are all equivalent up to a simple shift (3.14). For bi-Hamiltonian equations entirely different Hamiltonian formulations are known, which immediately give access to the integrability of these systems. All these operators can be used for the lifting procedure, the resulting systems differ only by a trivial "gauge" transformation, i.e., a shift of the new "gauge" variable π. This might give a simplifying point of view for considering multi-Hamiltonian formulations. Indeed, the compatibility conditions of Hamiltonian pairs as well as the essential hereditary property of recursion operators arise in an amazingly natural way in this context.

We have not discussed compatibility conditions of the proposed liftings w.r.t. to shifts of the gauge variable: μ_0 is fixed in all our considerations. Hence a natural next step would be the investigation of these lifts w.r.t. different Hamiltonian potentials $d\mu_0$. It can be shown that in the presence of mastersymmetries[15] certain covector fields are distinguished potentials of the Hamiltonian operators. The hierarchy of lifted systems considered here can in this case be enlarged by additional liftings of the original dynamical systems using these potentials for different Hamiltonian forms.

APPENDIX: THE LIFTS l_1 AND l_2

We give the definitions of the liftings l_1 (the "complete" lifts of Ref. 5) in a way that can also be applied to field equations such as the KdV equation. Let (u,v) be a local coordinate system on TM, then we define

(a) for $f \in F(M)$, $\quad l_1(f) := \langle df(u),v\rangle$;

(b) for $K \in X(M)$, $\quad l_1(K) := \begin{pmatrix} K(u) \\ K'(u)[v] \end{pmatrix}$;

(c) for $\gamma \in X^*(M)$, $\quad l_1(\gamma) := \begin{pmatrix} \gamma'(u)[v] \\ \gamma(u) \end{pmatrix}$;

(d) for $J \in T^0_2(M)$, $l_1(J) := \begin{pmatrix} J'(u)[v] & ; & J(u) \\ J(u) & ; & 0 \end{pmatrix}$;

(e) for $\Theta \in T^2_0(M)$, $l_1(\Theta) := \begin{pmatrix} 0 & ; & \Theta(u) \\ \Theta(u) & ; & \Theta'(u)[v] \end{pmatrix}$;

(f) for $\Phi \in T^1_1(M)$, $l_1(\Phi) := \begin{pmatrix} \Phi(u) & ; & 0 \\ \Phi'(u)[v] & ; & \Phi(u) \end{pmatrix}$.

All these definitions are invariant w.r.t. coordinate changes of the form $\tilde{u} = \tilde{u}(u)$, $\tilde{v} = (\partial \tilde{u}/\partial u)v$.

Let (u,π) be coordinates of T^*M. The liftings l_2 of Sec. IV are calculated as

(a) for $f \in F(M)$, $l_2(f;\Theta_0,\mu_0) := \langle df, \Theta_0(\pi - \mu_0) \rangle$;

(b) for $K \in X(M)$, $l_2(K;J_0,\mu_0) := \begin{pmatrix} K(u) \\ \widehat{K}(u,\pi) \end{pmatrix}$,

with $\widehat{K}(u,\pi) := -K'^*\pi + L_K\mu_0 + (L_K J_0) J_0^{-1}(\pi - \mu_0)$,

i.e., $\langle \widehat{K}, X \rangle = -\langle \pi, K'[X] \rangle + \langle L_K\mu_0, X \rangle + \langle (L_K J_0) J_0^{-1}(\pi - \mu_0), X \rangle$, $X \in T_u M$;

(c) for $\gamma \in X^*(M)$, $l_2(\gamma;\Theta_0,\mu_0) := \begin{pmatrix} \hat{\gamma}(u,\pi) \\ \Theta_0^*(u)\gamma(u) \end{pmatrix}$,

with $\hat{\gamma}(u,\pi) := \gamma'[\Theta_0\pi] + (\Theta_0'[\cdot]\pi)^*\gamma - L_{\Theta_0\mu_0}\gamma$,

i.e., $\langle \hat{\gamma}, X \rangle = \langle \gamma'[\Theta_0\pi], X \rangle + \langle \gamma, \Theta_0'[X]\pi \rangle - \langle L_{\Theta_0\mu_0}\gamma, X \rangle$, $X \in T_u M$;

(d) for $J \in T^0_2(M)$, $l_2(J;\Theta_0,\mu_0) := \begin{pmatrix} \hat{J}(u,\pi) & ; & J(u)\Theta_0(u) \\ \Theta_0^*(u)J(u) & ; & 0 \end{pmatrix}$,

with $\hat{J}(u,\pi) := J'[\Theta_0\pi] + J\Theta_0'[\cdot]\pi + (\Theta_0'[\cdot]\pi)^*J - L_{\Theta_0\mu_0}J$,

i.e., $\langle \hat{J}X_1, X_2 \rangle = \langle J'[\Theta_0\pi]X_1, X_2 \rangle + \langle J\Theta_0'[X_1]\pi, X_2 \rangle + \langle JX_1, \Theta_0'[X_2]\pi \rangle - \langle (L_{\Theta_0\mu_0}J)X_1, X_2 \rangle$, $X_1, X_2 \in T_u M$;

(e) for $\Theta \in T^2_0(M)$, $l_2(\Theta;J_0,\mu_0) := \begin{pmatrix} 0 & ; & \Theta(u)J_0^*(u) \\ J_0(u)\Theta(u) & ; & \widehat{\Theta}(u,\pi) \end{pmatrix}$,

with $\widehat{\Theta}(u,\pi) := J_0'[\Theta J_0^* \cdot]J_0^{-1}\pi + J_0\Theta'[J_0'\pi]J_0^* + J_0\Theta(J_0'[\cdot]J_0^{-1}\pi)^* - J_0(L_{J_0^{-1}\mu_0}\Theta)J_0^*$,

i.e., $\langle \widehat{\Theta}X_1, X_2 \rangle = \langle J_0'[\Theta J_0^* X_1]J_0^{-1}\pi, X_2 \rangle + \langle J_0\Theta'[J_0^{-1}\pi]J_0^* X_1, X_2 \rangle + \langle J_0'[\Theta^* J_0^* X_2]J_0^{-1}\pi, X_1 \rangle$
$- \langle J_0^* X_2, (L_{J_0^{-1}\mu_0}\Theta)J_0^* X_1 \rangle$, $X_1, X_2 \in T_u M$;

(f) for $\Phi \in T^1_1(M)$, $l_2(\Phi;J_0,\mu_0) := \begin{pmatrix} \Phi(u) & ; & 0 \\ \widehat{\Phi}(u,\pi) & ; & J_0(u)\Phi(u)J_0^{-1}(u) \end{pmatrix}$,

with $\widehat{\Phi}(u,\pi) := J_0'[\Phi \cdot]J_0^{-1}\pi + J_0\Phi'[J_0^{-1}\pi] - J_0\Phi J_0^{-1}J_0'[\cdot]J_0^{-1}\pi - J_0(L_{\Theta_0\mu_0}\Phi)$,

i.e., $\langle \widehat{\Phi}X_1, X_2 \rangle = \langle J_0'[\Phi X_1]J_0^{-1}\pi, X_2 \rangle + \langle J_0\Phi'[J_0^{-1}\pi]X_1, X_2 \rangle$
$- \langle J_0\Phi J_0^{-1}J_0'[X_1]J_0^{-1}\pi, X_2 \rangle - \langle J_0(L_{J_0^{-1}\mu_0}\Phi)X_1, X_2 \rangle$, $X_1, X_2 \in T_u M$.

All definitions are invariant w.r.t. cotangent bundle transformations (3.16).

[1] Y. Nutku, "Canonical formulation of shallow water waves," J. Phys. A **16**, 4195 (1983); "Hamiltonian formulation of the KdV equation," J. Math. Phys. **25**, 2007 (1984).

[2] P. J. Olver, "Dirac's theory of constraints in field theory and the canonical form of Hamiltonian differential operators," J. Math. Phys. **27**, 2495 (1986).

[3] D. D. Holm, J. E. Marsden, T. Ratiu, and A. Weinstein, "Nonlinear stability of fluid and plasma equilibria," Phys. Rep. **123**, 1 (1985).

[4] I. M. Gelfand and L. A. Dikii, "The resolvent and Hamiltonian pairs," Func. Anal. Appl. **11**, 93 (1977).

[5] K. Yano and S. Ishihara, *Tangent and Cotangent Bundles: Differential Geometry, Pure and Applied Mathematics,* Vol. 16 (Dekker, New York, 1973).

[6] P. J. Olver, "Evolution equations possessing infinitely many symmetries," J. Math. Phys. **18**, 1212 (1977).

[7] A. Weinstein, "The local structure of Poisson manifolds," J. Differ. Geom. **18**, 523 (1983).

[8] V. E. Zakharov and L. D. Faddeev, "Korteweg–de Vries equation: A completely integrable Hamiltonian system," Func. Anal. Appl. **5**, 280 (1972).

[9] T. Taniuti and K. Nishihara, *Nonlinear Waves, Monographs and Studies in Mathematics,* Vol. 15 (Pitman, Boston, 1983).

[10] B. Fuchssteiner and A. S. Fokas, "Symplectic structures, their Baecklund transformations, and hereditary symmetries," Physica D **4**, 47 (1981).

[11] Y. Choquet-Bruhat, C. de Witt-Morette, and M. Dillard-Bleick, *Analysis, Manifolds and Physics* (North-Holland, Amsterdam, 1977).

[12] B. Fuchssteiner, "Application of hereditary symmetries to nonlinear evolution equations," Nonlinear Anal. TMA **3**, 849 (1979).

[13] F. Magri, "A simple model of the integrable Hamiltonian equation," J. Math. Phys. **19**, 1204 (1968).

[14] I. M. Gelfand and I. Y. Dorfman, "Hamiltonian operators and algebraic structures related to them," Func. Anal. Appl. **13**, 248 (1979); "The Schouten bracket and Hamiltonian operators," **14**, 223 (1980).

[15] W. Oevel, "Mastersymmetries: weak action/angle structure for Hamiltonian and non-Hamiltonian systems," preprint 1986; "A geometrical approach to integrable systems admitting time dependent invariants," in *Topics in Soliton Theory and Exactly Solvable Nonlinear Equations*, edited by M. Ablowitz, B. Fuchssteiner, and M. Kruskal (World Scientific, Singapore, 1987).

Analytic solutions of the SU(3) Yang–Mills field equations with external sources

C. H. Oh and Rajesh R. Parwani
Department of Physics, Faculty of Science, National University of Singapore, Kent Ridge, Singapore 0511, Republic of Singapore

(Received 25 June 1987; accepted for publication 9 September 1987)

For external sources specified in the radial gauge frame, it is demonstrated how the SU(3) Yang–Mills field configuration can be constructed from the SU(2) solutions. This is explicitly illustrated for the type-II solutions that exhibit the bifurcation phenomenon.

I. INTRODUCTION

Since the work of Sikivie and Weiss,[1] there are by now many numerical as well as analytic solutions of the SU(2) Yang–Mills (YM) field equations with external sources.[2] For the gauge group SU(3), there are very few solutions available. In Ref. 3, Horvat and Viswanathan obtained numerical solutions for the response of the SU(3) YM field to an external spherical delta-shell source. Most recently Horvat discussed the stability problem of their SU(3) bifurcating solutions.[4] The purpose of the present paper is to indicate how analytic solutions for the SU(3) gauge group can be obtained from SU(2) solutions when the external source has a spherical symmetry and is specified in the radial gauge frame. We shall illustrate our method explicitly for the type-II solutions only as other solutions can be derived similarly. As in the SU(2) case, the SU(3) type-II solutions exhibit the bifurcation phenomenon. Bifurcation solutions in general correspond to a reduction of symmetry[5]; this is most easily seen in the case when the external source is specified in the Abelian gauge frame.[6] In passing, we note that bifurcation may play an important role in physics.[7]

II. THE GAUGE FIELD EQUATIONS AND SOLUTIONS

The Yang–Mills equations in the presence of an external static source $j(x)$ are given by

$$D_\mu F^{\mu\nu} = j^\nu \equiv \delta_0^\nu \rho. \tag{1}$$

For the SU(2) theory the radial ansatz[8]

$$A_0^a = n^a f(r)/(gr), \quad n^a = x^a/r, \tag{2a}$$

$$A_i^a = \varepsilon_{iaj} n^j [a(r) - 1]/(gr), \tag{2b}$$

$$\rho^a = n^a q(r)/g, \tag{2c}$$

simplifies Eq. (1) to a pair of coupled differential equations:

$$-a'' + (a^2 - 1)a/r^2 - af^2 r^2 = 0, \tag{3a}$$

$$-f'' + 2a^2 f/r^2 = rq. \tag{3b}$$

The prime means differentiation with respect to the argument. Similarly, for the SU(3) theory the spherically symmetric ansatz[3]

$$gA_{ab}^0 = i\varepsilon_{abc} n^c f_1(r)/r + (n_a n_b - \tfrac{1}{3}\delta_{ab}) f_2(r)/r, \tag{4a}$$

$$gA_{ab}^i = i(\delta_{ib} n_a - \delta_{ia} n_b)(G(r) - 1)/r, \tag{4b}$$

$$g\rho_{ab} = i\varepsilon_{abc} n_c q_1(r) + (n_a n_n - \tfrac{1}{3}\delta_{ab}) q_2(r) \tag{4c}$$

reduces Eq. (1) to the following coupled differential equations:

$$-G'' + (G^2 - 1)G/r^2 - (f_1^2 + f_2^2)G/r^2 = 0, \tag{5a}$$

$$-f_1'' + 2G^2 f_1/r^2 = 2rq_1, \tag{5b}$$

$$-f_2'' + 6G^2 f_2/r^2 = 2rq_2. \tag{5c}$$

As is noted in Ref. 3, by setting $q_2 = f_2 = 0$ the solutions of Eqs. (5) will also be solutions of Eqs. (3) with $a = G, f = f_1$, and $q = 2q_1$. We now show that any solution of the SU(2) equations (3) can be used to construct a solution for the SU(3) equation (5). Given a solution $a(r)$, $f(r)$ of Eq. (3), we define

$$G(r) = a(r), \tag{6a}$$

$$f_1(r) = f \sin w, \tag{6b}$$

$$f_2(r) = f \cos w, \tag{6c}$$

where the parameter w is restricted to $0 \leqslant w \leqslant \pi/2$. Equation (5a) then becomes identical to Eq. (3a) while Eqs. (5b) and (5c) define the source functions q_1 and q_2. We note that $q_1(r)$ is proportional to $q(r)$ for a fixed value of w.

The gauge-invariant energy of the solutions (5) is given by[3]

$$\xi = \frac{4\pi}{g} \int_0^\infty dr \left\{ \frac{1}{2}(f_1')^2 + \frac{1}{6}(f_2')^2 + (G')^2 \times \frac{1}{r^2}(f_1^2 + f_2^2)G^2 + \frac{1}{2r^2}(G^2 - 1)^2 \right\}. \tag{7}$$

The Casimir invariants for the SU(3) theory are[1]

$$C_1(r) = 2\,\mathrm{Tr}[\rho(r)]^2 = (4/g^2)(q_1^2 + \tfrac{1}{3}q_2^2), \tag{8a}$$

and

$$C_2(r) = 4\,\mathrm{Tr}[\rho(r)]^3 = (8/g^3)(\tfrac{1}{9}q_2^3 - q_1^2 q_2). \tag{8b}$$

These are used to define the gauge-invariant total external charges

$$Q = 4\pi \int_0^\infty r^2 [C_1(r)]^{1/2} dr = \frac{8\pi}{g} \int_0^\infty r^2 \left(q_1^2 + \frac{1}{3} q_2^2 \right)^{1/2} dr, \tag{9a}$$

and

$$R = 4\pi \int_0^\infty r^2 |C_2(r)|^{1/3} dr = \frac{8\pi}{g} \int_0^\infty r^2 \left| \frac{1}{9} q_2^3 - q_1^2 q_2 \right|^{1/3} dr. \tag{9b}$$

By comparing expressions (7) and (9) with the correspond-

0022-2488/88/010220-04$02.50

ing expressions for the SU(2) case,[9] it is easy to show that because of the relation (6), the finite energy–finite charge SU(2) solutions of (3) also render ξ [Eq. (7)], Q, and R finite. An example of such an analytic solution to (3) is given in Ref. 9:

$$a(r) = \tanh u, \tag{10a}$$

$$f(r) = [r^2(2u'^2 - u''/\tanh u) - 1]^{1/2} \operatorname{sech} u, \tag{10b}$$

with

$$u(r) = b(1/y - y^2), \quad y = r/c. \tag{10c}$$

The parameters b and c are positive. By using Eqs. (6), we find that the energy and charges Q and R for the corresponding SU(3) solutions can be written as

$$\xi = (1/c)H(b,w), \tag{11a}$$

$$Q = Q(b,w), \tag{11b}$$

$$R = R(b,w), \tag{11c}$$

where H, Q, and R are functions of b and w only.

III. BIFURCATION

In all subsequent discussions, we shall consider Eqs. (6), (10), and (11) for the particular case $w = \pi/4$ only. Numerical computation then shows that the minima of H, Q, and R occur, respectively, at $b = b_0 \equiv 0.5621$, $b = b_1 \equiv 0.6398$, and $b = b_2 \equiv 0.654\,49$. Also, in order for the gauge fields to be real we require that $b > 0.5410$.

In order to obtain bifurcating SU(3) solutions, we impose a relation[10] between b and c so that ξ and Q (or ξ and R) have their minima at the same values of the parameters. Consider first the bifurcation of ξ with Q. We parametrize c in terms of b as

$$c = mb + p. \tag{12}$$

Then using (11a) and (12), we see that $d\xi(b_1)/db = 0$ implies

$$p = m(1/k_1 - b_1), \tag{13}$$

where

$$k_1 = \frac{dH(b_1)/db}{H(b_1)} = 1.713.$$

Furthermore one can easily show that since $d^2H(b_1)/db^2 > 0$ then $d^2\xi(b_1)/db^2 > 0$ if and only if $m > 0$. The value of the energy at the bifurcation point is

$$\xi(b_1) = \frac{dH(b_1)/db}{m}. \tag{14}$$

Hence the bifurcation point depends on the value of m. Figure 1 shows a plot of ξ vs Q for $m = 1$. The corresponding solutions $a(r)$ and $f(r)$ together with the functions $q_1(r)$ and $q_2(r)$ near the bifurcation point are shown in Figs. 2–5, respectively. Note that $G(r) = a(r)$ and $f_1(r) = f_2(r) = f(r)/\sqrt{2}$; also $q_1(0) = q_1(\infty) = q_2(0) = q_2(\infty) = 0$. In Figs. 6 and 7, we have plotted the charge densities $\overline{Q}(r)$ and $\overline{R}(r)$, where

$$\overline{Q}(r) \equiv (q_1^2 + \tfrac{1}{3}q_2^2)^{1/2}, \tag{15a}$$

$$\overline{R}(r) \equiv |(\tfrac{1}{9}q_2^3 - q_1^2 q_2)|^{1/3}. \tag{15b}$$

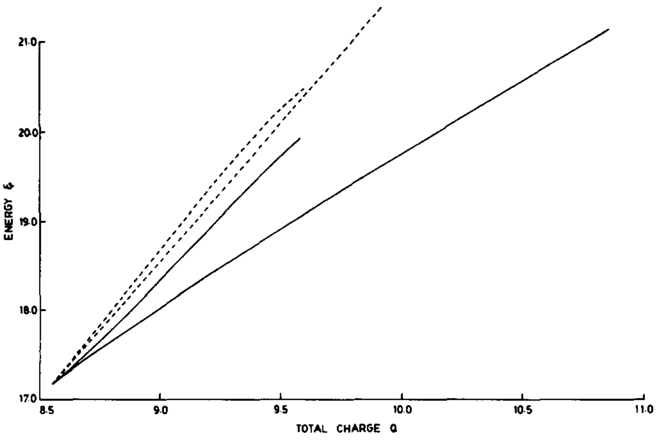

FIG. 1. Here ξ vs Q is shown for $m = 1$. The solid curve corresponds to the parametrization (12) while the dashed curve follows (16) with $\gamma = -1/k_1$.

Actually, one can construct infinitely many distinct pairs of branches emanating from each bifurcation point. For example, we set[10]

$$c = \alpha b^2 + \beta b + \gamma, \tag{16a}$$

with

$$\alpha = m/b_1 - (m/k_1 - \gamma)/b_1^2, \tag{16b}$$

$$\beta = 2(m/k_1 - \gamma)/b_1 - m, \tag{16c}$$

$$\gamma \leqslant m(1/k_1 - b_1). \tag{16d}$$

In Fig. 1, the dashed curve corresponds to $m = 1$ and $\gamma = -1/k_1$. The curves for $a(r)$, $f(r)$, etc., are similar to those of Figs. 2–7 and are not shown here.

To obtain the bifurcation of ξ with R, replace b_1 by b_2 and k_1 by $k_2 \equiv 1.804$ in Eqs. (13), (14), and (16). Figure 8 shows the bifurcation curves of ξ vs R for $m = 1$. The solid curve corresponds to $\gamma = 1/k_2 - b_2$ while the dashed curve corresponds to $\gamma = -1/k_2$. The curves for $a(r)$, $f(r)$, etc., are similar to those of Figs. 2–7.

IV. COMMENTS

(a) There are ways other than relation (6) of obtaining SU(3) solutions from the SU(2) ones. For example, we may set

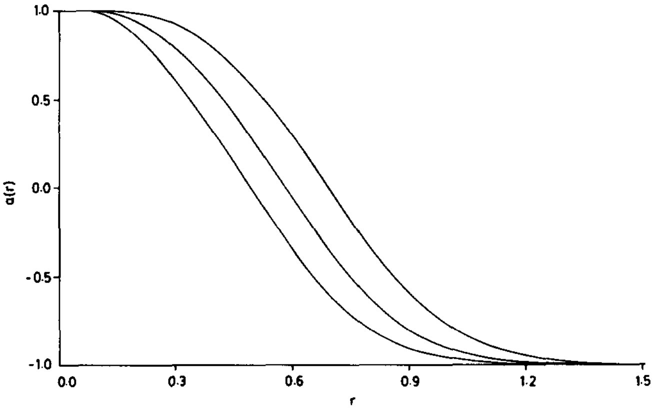

FIG. 2. Here $a(r)$ is shown for the solution (10) with $m = 1$ in (12). Starting from the curve with the smallest zero, these correspond to $b = 0.5500$, 0.6398, and 0.7500.

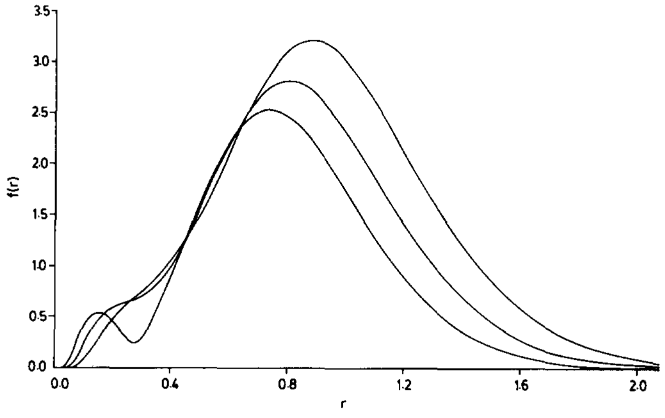

FIG. 3. Here $f(r)$ is shown corresponding to the $a(r)$ of Fig. 2. The curves are identified by noting that the peak value increases with b and $b = 0.5500$, 0.6398, 0.7500.

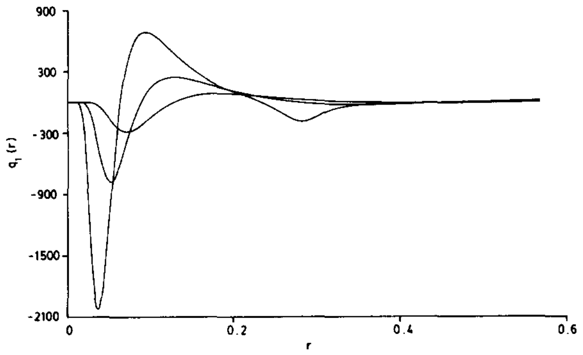

FIG. 4. Here $q_1(r)$ is shown for the $a(r)$ and $f(r)$ of Figs. 2 and 3. Starting from the curve with the most negative peak, these correspond to $b = 0.5500$, 0.6398, 0.7500.

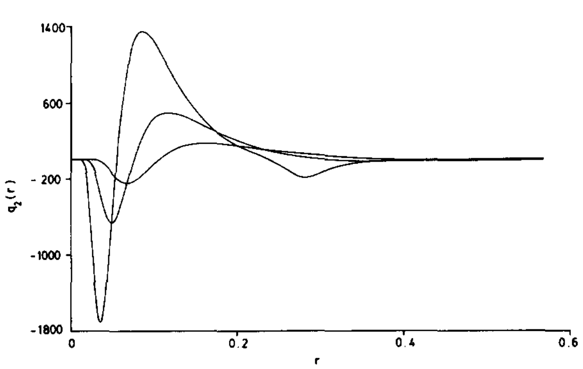

FIG. 5. Here q_2 is shown as for Fig. 4.

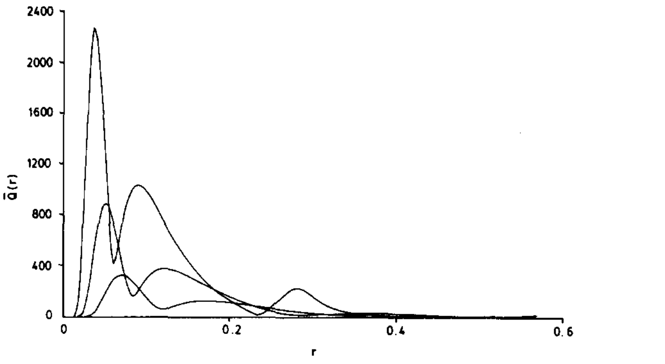

FIG. 6. The charge density $\overline{Q}(r)$ given by (15a) corresponding to the curves of Figs. 4 and 5. The peak value of these curves decreases as b increases; $b = 0.5500$, 0.6398, 0.7500.

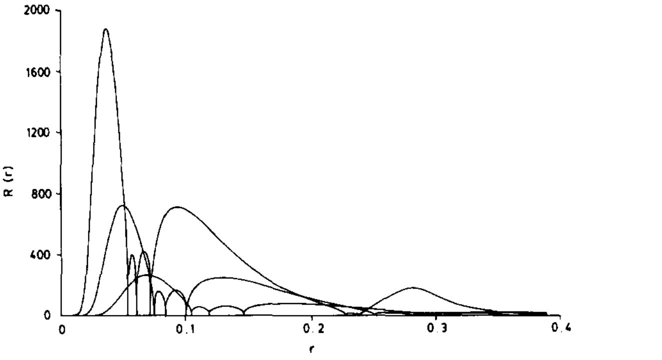

FIG. 7. The charge density $\overline{R}(r)$ given by (15b) corresponding to the curves of Figs. 4 and 5. The peak values of these curves decreases as b increases; $b = 0.5500$, 0.6398, 0.7500.

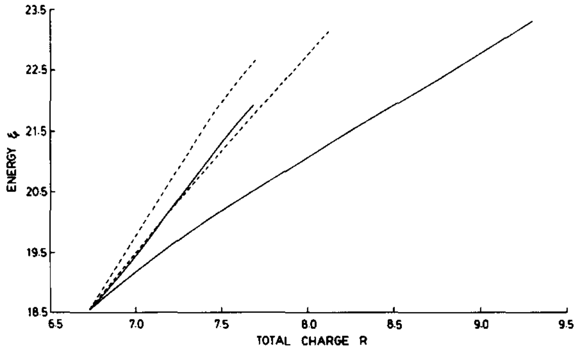

FIG. 8. Here ξ vs R is shown for $m = 1$. The curves correspond to the parametrization (16) when b_1 and k_1 are replaced by b_2 and k_2. The solid curve has $\gamma = 1/k_2 - b_2$ while for the dashed curve $\gamma = -1/k_2$.

$$G(r) = a(r), \quad f_1(r) = f\exp(-Br^n),$$
$$f_2(r) = f[1 - \exp(-2Br^n)]^{1/2},$$

with the parameters B and n being positive.

(b) Our definition of the charges in Eqs. (9) does not correspond to that used by the authors of Ref. 3. The gauge-dependent charges that would correspond to their usage are

$$S_1 = \frac{4\pi}{g}\int_0^\infty r^2 q_1\, dr$$

and

$$S_2 = \frac{4\pi}{g}\int_0^\infty r^2 q_2\, dr.$$

Using Eqs. (5) and (6) and because of the boundary conditions on finite energy–finite charge solutions, we have

$$S_2 = 3S_1(\cot w).$$

(c) Our SU(3) solution as given by (6) and (10) is a type-II solution[3,8] [i.e., $G(\infty) = -1$]. Following Ref. 3, the solution for $w = 0$ belongs to group 1, that for $w = \pi/2$ to group 2, while those for $0 < w < \pi/2$ belong to group 3.

(d) Unlike Ref. 3, here both charges Q and R vary as we observe the bifurcation of ξ with one of them.

(e) Following Ref. 10, the bifurcating solutions obtained in the last section are weakly bifurcating since the corresponding points on the two branches arise from different charge densities although their total charges are the same.

(f) As in Ref. 3, we set $j_i(x) = 0$ in Eq. (1) since for nonvanishing external current density the total energy becomes gauge dependent.

[1] P. Sikivie and N. Weiss, Phys. Rev. Lett. **40**, 1411 (1978); Phys. Rev. D **18**, 3809 (1978).

[2] For a review and references, see H. Arodz, Acta Phys. Polon. B **14**, 825 (1983).

[3] D. Horvat and K. S. Viswanathan, Phys. Rev. D **23**, 937 (1981).

[4] D. Horvat, Phys. Rev. D **34**, 1197 (1986).

[5] G. Cicogna, Lett. Nuovo Cimento **31**, 600 (1981).

[6] C. H. Oh, S. N. Chow, and C. H. Lai, Phys. Rev. D **39**, 1334 (1984).

[7] N. P. Chang and L. N. Chang, Phys. Rev. Lett. **54**, 2407 (1985); E. Malec, J. Math. Phys. **23**, 21 (1982).

[8] R. Jackiw, L. Jacobs, and C. Rebbi, Phys. Rev. D **20**, 474 (1979); R. Jackiw and P. Rossi, *ibid.* **21**, 426 (1980).

[9] C. H. Oh, R. Teh, and W. K. Koo, Phys. Rev. D **25**, 3263 (1982).

[10] C. H. Oh and R. R. Parwani, Phys. Rev. D **36**, 2527 (1987).

Duality and orthogonal transitivity in dimensional reduction

Brandon Carter
Institute for Theoretical Physics, University of California at Santa Barbara, Santa Barbara, California 93106 and Groupe d'Astrophysique Relativiste, Centre National de la Recherche Scientifique, D. A. R. C., Observatoire de Paris, 92195 Meudon, France

(Received 3 September 1987; accepted for publication 9 September 1987)

Conditions under which a symmetry group action can be expected to be orthogonally transitive are investigated within the generalized Kaluza–Klein framework in terms of the dimensional reduction scheme whereby a higher-dimensional space is regarded as a bundle over a lower-dimensional base space formed as its quotient with respect to the surfaces over which the group action is transitive, so that orthogonal transitivity is interpretable as meaning the existence of a section that is everywhere orthogonal to the fibers. It is shown that if a certain comoving source condition is satisfied locally, then the system admits a dual reformulation whereby trivectors are introduced in place of the usual gauge-potential covectors, and that under these circumstances (but not otherwise) orthogonal transitivity will hold subject to suitable global boundary conditions provided appropriate local signature inequalities are satisfied everywhere, the simplest (but not the only) possibility being that in which both the metric and the relevant coupling matrix are positive definite.

I. INTRODUCTION

The purpose of this work is to use the methods of dimensional reduction in Kaluza–Klein type (generalized Einstein or Einstein–Maxwell) spaces, to make a systematic study of conditions under which one can expect the action of any symmetry group that may be present to be *orthogonally transitive*. Within the conceptual framework of dimensional reduction, in which the higher-dimensional space is considered as a bundle over a lower-dimensional base space formed as its quotient with respect to fibers that are the surfaces of transitivity of the symmetry group action, the condition of orthogonal transitivity means the existence of a bundle section (and hence of a space-filling congruence of such sections) that is everywhere orthogonal to the fibers. This means that it is possible to use a local reference system in which the metric has no cross components between a class of coordinates that is comoving in the sense of being constant over each surface of transitivity, and a complementary class (that can be taken to be ignorable in the Abelian case to be studied here) that are constant on each orthogonal section.

Orthogonal transitivity (which includes staticity as a special case for a one-parameter group) is often an appreciable simplification, which has frequently been postulated for convenience, but without mathematical justification at the time, by many workers in diverse contexts, in the hope either that the conclusions drawn thereby would remain qualitatively valid in more general nonorthogonally transitive cases or else that subsequent work would establish that orthogonal transitivity should necessarily hold in any case. A classic example, which provided much of the underlying motivation for the present work, is the postulate of staticity by Israel in his historic investigations[1,2] of the extent to which the Schwarzschild black hole is unique. It was the subsequent study of black hole equilibrium states under more general conditions, and in particular the nonstatic stationary-axisymmetric case first dealt with by Papapetrou[3] that motivated a first systematic investigation[4] (in effect a prototype of the more extended investigation carried out here) of the question of the necessity of orthogonal transitivity under more general circumstances. The particular problem raised by the postulate of statisticity in Israel's work, at least in the pure vacuum case,[1] was partially solved by the adaptation to the black hole boundary conditions by Hawking[5] of an earlier result due to Lichnerowicz.[6] General reviews of questions of orthogonal transitivity in the context of black-hole equilibrium state boundary condition problems have been given by the present author,[7,8] including a generalization of the Hawking–Lichnerowicz theorem to cover cases involving the presence of an electromagnetic field, a sign error in the original version[7] having been corrected in the latter.[8] The present investigation grew out of an alternative approach to the same problem,[9] whereby instead of the essentially covariant treatment used in the reviews just cited, a procedure based on the use of quantities defined with respect to the bundle structure and the quotient space of dimensional reduction was used, following lines suggested by work of Breitenlohner, Maison, and Gibbons.[10]

The widespread current popularity of higher-dimensional theories of diverse types suggests the potential utility of the present investigation, which shows how techniques originally developed in a specialized four-dimensional context can be generalized to situations in which arbitrary numbers of dimensions are involved, thereby making available large classes of orthogonal transitivity theorems. These results can be applied to diverse combinations of (induced or projected) signature conditions, but are dependent on the verification of the appropriate boundary conditions (which did turn out to be satisfied in the black-hole applications mentioned above, but which must be scrupulously checked each time a new application is envisaged). The results in question all depend on the invocation of a requirement that any material sources must (in a sense to be precisely defined) be *comoving* with the surfaces of transitivity (if not, orthogonal transitivity would be violated locally, even in the presence of favorable boundary conditions). As an interme-

0022-2488/88/010224-11$02.50

diate step, it is shown that (independently of any boundary conditions) this comoving source requirement is locally sufficient to ensure the existence of a useful dual reformulation whereby trivectors are introduced instead of the usual gauge potential covectors.

From the point of view of a physical interpretation in terms of fields induced by a higher-dimensional Kaluza–Klein type projection process on an intermediate-dimensional "physical" space with a symmetry group whose action determines a lower-dimensional quotient space, the most interesting suitable combination of signature conditions is that for which the field coupling matrix and the projected metric on the lower-dimensional space are both (let us say) *positive* definite, but for which the higher-dimensionally induced metric (as distinct from the physical projected metric) on the surfaces of transitivity of the group action is of opposite, i.e. (let us say), *negative* definite signature. The most obvious alternative possible combination of conditions is of course simply that for which the overall metric on the higher-dimensional space itself (and hence automatically each of the intermediate and lower-dimensional induced or projected metrics, including the coupling matrix) is of (let us say) positive definite signature.

II. BASIC VARIATIONAL FORMULATION OF THE REDUCIBLE SYSTEMS UNDER STUDY

The subject of investigation in this and the following sections will be the class of systems specifiable by a variation principle on a manifold of let us say m dimensions with local coordinates x^μ, $\mu = (1,...,m)$, in terms of a Lagrangian scalar density $\mathscr{L}$ of the form

$$\mathscr{L} = \mathscr{L}_{\text{geom}} + \mathscr{L}_{\text{vect}} + \mathscr{L}_{\text{matt}} \tag{2.1}$$

in which the relevant field variables are taken to be a set of $(m+q)$ independent one-forms, $\boldsymbol{\theta}^{\hat\alpha}$, $\mathbf{A}^X$ for $\alpha = (1,...,m)$, $X = (m+1,...,m+q)$, with components $\theta_\mu{}^{\hat\alpha}$, $A_\mu{}^X$, of which the former constitute a Cartan frame determining a (pseudo)-metric tensor

$$g_{\mu\rho} = g_{\hat\alpha\hat\gamma}\theta_\mu{}^{\hat\alpha}\theta_\rho{}^{\hat\gamma}, \tag{2.2}$$

where $g_{\hat\alpha\hat\gamma}$ is a *constant* diagonal matrix, with component values normalized to ± 1, and where the $\mathbf{A}^X$ are potentials for q (one or several) Maxwell-type fields of the form

$$F_{\mu\nu}{}^X = 2\partial_{[\mu}A_{\nu]}{}^X \tag{2.3}$$

(using square brackets to denote antisymmetrization, and ∂_μ to denote partial differentiation with respect to x^μ), the corresponding Lagrangian densities having the standard (respectively, Einstein–Hilbert and Maxwellian type) forms

$$\mathscr{L}_{\text{geom}} = (-\|\theta\|/16\pi G)R \tag{2.4}$$

and

$$\mathscr{L}_{\text{vect}} = (-\|\theta\|/16\pi G)G_{XY}F_{\mu\nu}{}^X F^{\mu\nu Y}, \tag{2.5}$$

where $|\theta|$ denotes the determinant of the obligatorily nonsingular $(m\times m)$ frame matrix $\theta_\mu{}^{\hat\alpha}$ so that its modulus $\|\theta\|$ is just the naturally associated measure density, and where R in (2.4) denotes the corresponding Ricci scalar as determined by the metric $g_{\mu\nu}$, which is of course also used in the standard way for the index raising involved in (2.5), while, finally, G is just Newton's constant and the coupling matrix G_{XY} might either be fixed or else might depend on independent matter fields. The behavior of any such (e.g., scalar or perfect fluid type) matter fields as may be present will be governed by the remaining Lagrangian contribution $\mathscr{L}_{\text{matt}}$ whose detailed form will not concern us here except insofar as it determines the relevant source current and energy-momentum densities, with components $\mathscr{J}_X{}^\mu$ and $\mathscr{T}_{\hat\alpha}{}^\mu$ defined by

$$\mathscr{J}_X{}^\mu = \delta\mathscr{L}_{\text{matt}}/\delta A_\mu{}^X \tag{2.6}$$

and

$$\mathscr{T}_{\hat\alpha}{}^\mu = \delta\mathscr{L}_{\text{matt}}/\delta\theta_\mu{}^{\hat\alpha}. \tag{2.7}$$

If the matter is bosonic so that the frame forms enter only in the metric combination (2.2), the latter will be replaceable by

$$\mathscr{T}^{\mu\nu} = 2(\delta\mathscr{L}_{\text{matt}}/\delta g_{\mu\nu}). \tag{2.8}$$

(The more traditional forms of the energy momentum tensor $T^{\mu\nu}$ and current vectors $j_X{}^\mu$ associated with the matter will be obtainable if required as $T^{\mu\nu} = \|\theta\|^{-}\mathscr{T}_{\mu\nu}$ and $j_X{}^\mu = \|\theta\|^{-1}\mathscr{J}_X{}^\mu$.) The Lagrangian density $\mathscr{L}_{\text{matt}}$ may include a contribution governing the coupling parameters G_{XY} as independent fields, or alternatively by the G_{XY} may have the status of a fixed background as in the standard Einstein–Maxwell theory which is obtained by taking $q = 1$ with the single coupling element having the constant value $G_{m+1,m+1} = G$ (where G is Newton's constant and $m = 4$ in ordinary space-time).

Membership of the class of theoretical models set up in the preceding paragraph is characterized (except in the particular two-dimensional case which needs a slightly modified special treatment as described below) of being preservable, in the manner to be described in the next section, under the effect of *dimensional reduction* onto the quotient space defined with respect to the action of any continuous symmetry group actions that might be manifested by the ignorability of a suitably chosen subset of say p independent coordinates, x^r, $r = n+1,...,m$ where $n = m - p$. This ("reducibility") property has been well known since its early exploitation (in the construction of higher-dimensional theories) by Kaluza[11] and Klein[12] and has more recently been shown by DeWitt,[13] Kerner,[14] Cho,[15] and others[16–18] to be generalized to dimensional reduction with respect to noncommuting symmetry group actions (i.e., ones that cannot be made manifest by any choice of simultaneously ignorable coordinates) by allowing the Maxwellian type of field considered here to be generalized to the wider class of Yang–Mills type fields for non-Abelian groups. Many of the ideas in the present paper could also be extended to apply to the non-Abelian case, but to avoid distraction from the essential points we shall restrict our attention in the present work to the technically simpler commuting case, which is more than sufficient to cover all the applications[8,9] referred to in the Introduction, namely stationary black holes in Einstein–Maxwell theory in which the base (quotient) space dimension is $n = 3$ or (in the stationary axisymmetric case) $n = 2$, and the total dimension is $m = 4$ or (in the Kaluza–Klein treatment) $m = 5$.

For our present purposes (and indeed for many others) the actual evaluation of the curvature is most conveniently carried out by the Cartan method, as based on the use of the frame connection of one-forms with mixed components $\gamma_\mu{}^{\hat\alpha}{}_{\hat\beta}$, as defined by the condition that the frame-covariant derivative of the frame forms should vanish, i.e.,

$$\nabla_\mu \theta_\rho{}^{\hat\alpha} + \gamma_\mu{}^{\hat\alpha}{}_{\hat\gamma}\theta_\rho{}^{\hat\gamma} \tag{2.9}$$

together with the antisymmetry condition

$$\gamma_{\mu(\hat\alpha\hat\beta)} = 0 \tag{2.10}$$

(using parentheses for symmetrization) ensuring the vanishing of the gauge covariant derivative of the matrix $g_{\hat\alpha\hat\beta}$ used for lowering of the frame indices (which are systematically distinguished from coordinate indices by the use of a hat). Using Cartan's trick of antisymmetrizing so as to be able to replace covariant differentiation (with respect to the metric $g_{\mu\nu}$) by simple partial differentiation, we use (2.9) to evaluate the exterior derivatives $\underline{S}^{\hat\alpha} = \partial \wedge \underline{\theta}^{\hat\alpha}$ of the frame forms, with components given by

$$S_{\mu\rho} = 2\partial_{[\mu}\theta_{\rho]}{}^{\hat\alpha}, \tag{2.11}$$

thereby obtaining the relation $\underline{S}^{\hat\alpha} = -\underline{\gamma}^{\hat\alpha}{}_{\hat\beta} \wedge \underline{\theta}^{\hat\beta}$, or in component notation

$$2\theta_{[\mu}{}^{\hat\gamma}\gamma_{\rho]}{}^{\hat\alpha}{}_{\hat\gamma} = S_{\mu\rho}{}^{\hat\alpha} \tag{2.12}$$

which may be solved conjointly with (2.10) to give the connection components explicitly in the form

$$\gamma_{\hat\alpha\hat\beta\hat\gamma} = \tfrac{1}{2}(S_{\hat\alpha\hat\beta\hat\gamma} + S_{\hat\gamma\hat\alpha\hat\beta} - S_{\hat\beta\hat\gamma\hat\alpha}), \tag{2.13}$$

where contractions with $\theta_\mu{}^{\hat\alpha}$ or with the corresponding contravariant frame matrix $\theta_{\hat\alpha}{}^\mu$ as defined by

$$\theta_{\hat\alpha}{}^\mu\theta_\mu{}^{\hat\gamma} = \delta^{\hat\gamma}_{\hat\alpha} \tag{2.14}$$

are used for conversion from coordinate to frame indices and vice versa.

Starting from the standard Cartan expression

$$R_{\mu\rho}{}^{\hat\alpha}{}_{\hat\gamma} = 2\partial_{[\mu}\gamma_{\rho]}{}^{\hat\alpha}{}_{\hat\gamma} + 2\gamma_{[\mu}{}^{\hat\alpha}{}_{|\hat\beta|}\gamma_{\rho]}{}^{\hat\beta}{}_{\hat\gamma} \tag{2.15}$$

for the mixed components of the Rieman tensor, one sees that its Ricci contraction

$$R = R_\mu{}^\mu, \quad R_{\mu\nu} = R_{\rho\mu}{}^\rho{}_\nu \tag{2.16}$$

will have the convenient tensorial expression

$$R = \gamma^{\mu\nu\rho}\gamma_{\nu\rho\mu} + \gamma_\rho{}^{\rho\mu}\gamma_\nu{}^\nu{}_\mu - 2\nabla_\mu\gamma_\rho{}^{\rho\mu}, \tag{2.17}$$

where the final divergence term may of course be ignored insofar as the application of the variational principle to (2.4) is concerned.

III. FORMULATION OF THE REDUCTION

We now consider the situation in which all the fields under consideration are invariant under the action of a set of let us say p (Killing) vector fields $\mathbf{k}_r$, with components $k_r{}^\mu$, that mutually commute so that by introducing comoving coordinates x^i for $i = 1,\ldots,n$, with $n = m - p$ it is possible to choose the remaining coordinates x^r, $r = n + 1,\ldots,m$, so as to be simultaneously ignorable, i.e., so that we have

$$\partial_r = 0 \quad (r = n + 1,\ldots,m) \tag{3.1}$$

for any of the fields under consideration (which means that the Killing vector fields may be taken to be specified by $k_r{}^\mu = \delta^\mu_r$).

Under these circumstances the original m-dimensional space, $\mathscr{M}$ say, can be conveniently envisaged as a bundle with p-dimensional fibers over an n-dimensional quotient (base) space, ${}^\flat\mathscr{M}$ say, with corresponding induced coordinates x^i ($i = 1,\ldots,n$), and the system can be reformulated in a natural way in terms of "reduced" fields over the base space, which we shall systematically distinguish from their higher-dimensional analogs by prefixing the musical symbol ${}^\flat$. This notation will be part of a systematic scheme in which the prefix ${}^\sharp$ is reserved for corresponding "augmented" field quantities in a higher-dimensional Kaluza–Klein type extended manifold to be described in Sec. V, while the prefix ${}^\natural$ will be used to distinguish the natural vertical (p-dimensional fiber) restrictions of fields from their full (m-dimensional) analogs.

A very convenient recent description of the appropriate dimensional reduction formalism has been given by Scherk and Schwarz.[19] One starts by choosing the frame vectors $\boldsymbol{\theta}_{\hat\alpha}$ with components $\theta_{\hat\alpha}{}^\mu$ (given in accordance with the usual convention by the matrix inverse to the corresponding one-form component matrix $\theta_\mu{}^{\hat\alpha}$) in such a way as to simplify the structure as much as possible by taking the last p of them to lie in the tangent subspace spanned by the Killing vectors $\mathbf{k}_r$, which means that we shall have

$$\theta_{\hat m}{}^i = 0 \quad (i = 1,\ldots,n;\ \hat m = n + 1,\ldots,m). \tag{3.2}$$

It then follows that the corresponding one-forms $\underline{\theta}^{\hat\alpha}$ will be characterized by

$$\theta_r{}^{\hat a} = 0 \quad (\hat a = 1,\ldots,n;\ r = n + 1,\ldots,m), \tag{3.3}$$

where we adopt a convention of using early Roman letters ($\hat a,\hat b,\hat c,\hat d$ for frame indices and h,i,j,k for base-space coordinate indices) running over the base range, $1,\ldots,n$, and to use late Roman letters [$\hat m,\hat n,\hat p,\hat q$ for frame indices and r,s,t,u for fiber (ignorable) coordinate indices] running over the complementary range $n + 1,\ldots,m$. In order for the reduction to be nonsingular, one must exclude the possibility of the surfaces of transitivity of the symmetry action (i.e., the fibers) being null, which is equivalent to the requirement that the p-dimensional determinant, $|{}^\natural\theta|$ say, of the purely vertical, i.e., fiber restricted part $\theta_r{}^{\hat m}$ of the frame matrix be everywhere nonzero in the space region under consideration. Under these circumstances the ($p\times p$) components of the "natural," (i.e., fiber restricted) inverse of the fiber frame with components $\theta^r_{\hat m}$ will coincide with the corresponding components of the full ($m\times m$) frame, i.e., we shall have

$$\theta_s{}^{\hat m}\theta_{\hat m}{}^r = \delta^r_s. \tag{3.4}$$

It now follows that we may introduce a set of p base-space one-forms $\underline{\alpha}^r$ (collectively interpretable as a single fiber vector valued one-form) with mixed components α^r_i by the defining condition that after the imposition of (3.2) and (3.3) the remaining cross components (between fiber and base) of the frame one-forms should be expressible in the form

$$\theta_i{}^{\hat m} = 2\alpha_i{}^r\theta_r{}^{\hat m}, \tag{3.5}$$

and hence that the remaining cross components of the inverse frame vectors should be given by

$$\theta_{\hat a}{}^{r} = -2\theta_{\hat a}{}^{i}\alpha_{i}{}^{r}. \tag{3.6}$$

After the specification of the fiber components $\theta_r{}^{\hat m}$ (which may be considered as a set of p^2 base-space scalars) and of the components $\alpha_i{}^r$ (which may be considered as representing a set of p base space one-forms $\underline{\alpha}^r$) the specification of the full m-dimensional frame $\theta_\mu{}^{\hat a}$ may be completed by the specification of a base frame ${}^\flat\theta_i{}^{\hat a}$ (which may be considered as representing a further set of n base space one-forms ${}^\flat\underline{\theta}^{\hat a}$) which determine the corresponding components of the full m-dimensional frame by an expression of the form

$$\theta_i{}^{\hat a} = \sigma\,{}^\flat\theta_i{}^{\hat a}, \tag{3.7}$$

or equivalently

$$\theta_{\hat a}{}^{i} = \sigma^{-1}\,{}^\flat\theta_{\hat a}{}^{i}, \tag{3.8}$$

where σ is a conformal scalar field whose choice will not be specified until later on.

Having thus expressed the frame forms $\underline{\theta}^{\hat a}$ in terms of corresponding base space scalars and one-forms, we now do the same for the field one-forms $\underline{A}^{x}$, defining base-space scalar potentials Φ_r^x and base-space one-forms ${}^\flat\underline{A}^{x}$ by

$$A_i{}^X = {}^\flat A_i^{\ X} + 2\alpha_i{}^r\Phi_r{}^X, \tag{3.9}$$

$$A_r{}^X = \Phi_r{}^X, \tag{3.10}$$

which are equivalent to

$$A_{\hat a}{}^X = \sigma^{-1}\,{}^\flat A_{\hat a}{}^X, \quad A_{\hat m}{}^X = \theta_{\hat m}{}^r\Phi_r{}^X. \tag{3.11}$$

The foregoing reduction implies that, in terms of the base-space metric

$${}^\flat g_{ij} = g_{\hat m\hat n}\,{}^\flat\theta_i{}^{\hat m}\,{}^\flat\theta_j{}^{\hat n}, \tag{3.12}$$

the full m-dimensional metric will be expressible by

$$g_{\mu\nu}\,dx^\mu\,dx^\nu = \sigma^2\,{}^\flat g_{ij}\,dx^i\,dx^j + g_{rs}(dx^r + 2\alpha_i{}^r\,dx^i)(dx^s + 2\alpha_j{}^s\,dx^j), \tag{3.13}$$

while the corresponding expression for the vector fields will be

$$A_\mu{}^X\,dx^\mu = {}^\flat A_i{}^X\,dx^i + \Phi_r{}^X(dx^r + 2\alpha_i{}^r\,dx^i). \tag{3.14}$$

To express the full field two-form $\underline{F}$ in terms of corresponding reduced (base-space) field quantities, we introduce base space one-forms and two-forms of, respectively, "electric" and "magnetic" type by the definitions

$$E_{ir}{}^X = \partial_i\Phi_r{}^X, \quad B_{ij}{}^X = {}^\flat F_{ij}{}^X + 2w_{ij}{}^r\Phi_r{}^X, \tag{3.15}$$

where the base-space field two-forms ${}^\flat\mathbf{F}$ are defined as the exterior products

$${}^\flat F_{ij}{}^X = 2\,\partial_{[i}{}^\flat A_{j]}{}^X \tag{3.16}$$

and the geometric "twist" two-forms $\boldsymbol{\omega}^r$ are defined analogously as the exterior products

$$\omega_{ij}{}^r = 2\,\partial_{[i}\alpha_{j]}{}^r. \tag{3.17}$$

In terms of the quantities introduced in this way, the frame components of the field two-forms will be given simply by

$$F_{\hat a\hat b}{}^X = \sigma^{-2}B_{\hat a\hat b}{}^X, \quad F_{\hat m\hat a}{}^X = \sigma^{-1}E_{\hat m\hat a}{}^X, \quad F_{\hat m\hat n}{}^X = 0. \tag{3.18}$$

The analogous expressions for the frame components of the connection (which we need in order to evaluate the Ricci scalar) can also be obtained [from (2.11) and (2.13)] in terms of the same base-space quantities that have just been introduced above, in the form

$$\begin{aligned}
\gamma_{\hat a\hat b\hat c} &= \sigma^{-1}\,{}^\flat\gamma_{\hat a\hat b\hat c} + 2\sigma^{-2}g_{\hat a[\hat b}\,{}^\flat\theta_{\hat c]}{}^i\,\partial_i\sigma,\\
\gamma_{\hat a\hat b\hat m} &= \gamma_{\hat m\hat b\hat a} = \sigma^{-2}\theta_{\hat m r}\omega_{\hat a\hat b}{}^r,\\
\gamma_{\hat a\hat m\hat n} &= \sigma^{-1}\,{}^\flat\theta_{\hat a}{}^i\theta_{[\hat m|}{}^r\,\partial_i\theta_{|\hat n]r},\\
\gamma_{\hat m\hat n\hat a} &= \tfrac{1}{2}\sigma^{-1}\theta_{\hat m}{}^r\theta_{\hat n}{}^s\,{}^\flat\theta_{\hat a}{}^i\,\partial_i g_{rs}, \quad \gamma_{\hat m\hat n\hat p} = 0,
\end{aligned} \tag{3.19}$$

from which it can be seen that the contraction appearing in the expression for the Ricci scalar will be given by

$$\gamma_{\hat\gamma}{}^{\hat\gamma\hat a} = \sigma^{-1}\{{}^\flat\gamma_{\hat c}{}^{\hat c\hat a} + {}^\flat\nabla^{\hat a}\ln(\sigma^{n-1}\|{}^\natural\theta\|)\}, \quad \gamma_{\hat\gamma}{}^{\hat\gamma\hat m} = 0. \tag{3.20}$$

Since we can express the full m-dimensional volume-measure density, $\|\theta\| = \|g\|^{1/2}$, in terms of the separate (n-dimensional) base-space (p-dimensional) fiber-space measures, respectively $\|{}^\flat\theta\| = \|{}^\flat g\|^{1/2}$ and $\|{}^\natural\theta\| = \|{}^\natural g\|^{1/2}$, in the form

$$\|\theta\| = \sigma^n\|{}^\flat\theta\|\,\|{}^\natural\theta\|, \tag{3.21}$$

it can be seen that it will be convenient to introduce the abbreviation

$$\rho = \sigma^{n-2}\|{}^\natural\theta\| \tag{3.22}$$

in order to obtain a simple expression relating the measure-weighted Ricci density, $\|\theta\|R$ for the full m-dimensional space to the corresponding reduced measure-weighted Ricci density, $\|{}^\flat\theta\|{}^\flat R$ for the n-dimensional base space. Transferring an unwanted divergence term to the left-hand side, one obtains the basic general purpose formula

$$\begin{aligned}
&\|\theta\|R + 2\|{}^\flat\theta\|{}^\flat\nabla_i\{\rho\,{}^\flat\nabla^i\ln(\rho\sigma)\}\\
&\quad = \rho\|{}^\flat\theta\|{}^\flat R + \|{}^\flat\theta\|({}^\flat\nabla_i\rho){}^\flat\nabla^i\ln(\rho\sigma^2)\\
&\qquad + \rho\|{}^\flat\theta\|\{(2-n)({}^\flat\nabla_i\ln\sigma){}^\flat\nabla^i\ln\sigma\\
&\qquad + \tfrac{1}{4}({}^\flat\nabla_i g_{rs}){}^\flat\nabla^i\,{}^\natural g^{rs} + \sigma^{-2}g_{rs}\omega_{ij}{}^r\omega^{ijs}\},
\end{aligned} \tag{3.23}$$

the corresponding formula for the generalized Maxwell contribution being

$$\begin{aligned}
&\|\theta\|G_{XY}F_{\mu\nu}{}^X F^{\mu\nu Y}\\
&\quad = \rho\|{}^\flat\theta\|G_{XY}(\sigma^{-2}B_{ij}{}^X B^{ijY} + 2\,{}^\natural g^{rs}E_{ri}^{\ X}E_s^{\ iY}),
\end{aligned} \tag{3.24}$$

where in accordance with the notation system introduced at the beginning of this section, ${}^\natural g^{rs}$ denotes the components of the inverse of the induced metric on the fibers, i.e., the natural matrix inverse of g_{rs}, which is not to be confused with the corresponding components of the full inverse metric $g^{\mu\nu}$ to which it can be seen [by (3.4) and (3.6)] to be related by

$$g^{rs} = {}^\natural g^{rs} + 4\sigma^{-2}\,{}^\flat g^{ij}\alpha_i{}^r\alpha_j{}^s. \tag{3.25}$$

In the case where the base-space dimension n is only 2, we shall have a special situation,

$$n = 2 \Rightarrow \rho = \|{}^\natural\theta\|. \tag{3.26}$$

Under these conditions the standard way of choosing the as yet unspecified conformal factor σ is to exploit the possibility of conformally adjusting the two-dimensional Ricci curvature to zero,

$$ {}^\flat R = 0, \tag{3.27}$$

thereby eliminating the "geometric" contribution, so that the preceding expression (3.23) simplifies to

$$\begin{aligned}\|\theta\|R + \|{}^\flat\theta\|\,{}^\flat\nabla_i\{\rho\,{}^\flat\nabla^i\ln(\rho^2\sigma^2)\} &= \|{}^\flat\theta\|({}^\flat\nabla_i\rho)\,{}^\flat\nabla^i\ln(\rho\sigma^2) + \rho\|{}^\flat\theta\| \\ &\quad\times\{\tfrac14\,{}^\flat\nabla_i g_{rs}\,{}^\flat\nabla^i\,{}^\natural g^{rs} + \sigma^{-2}g_{rs}\omega_{ij}{}^r\omega^{ijs}\}.\end{aligned}\tag{3.28}$$

We shall mainly be concerned here with the general case, with base-space dimension $n\geqslant3$, for which the geometric curvature cannot be conformally eliminated, but for which one can instead convert its contribution to the standard Einstein–Hilbert form $\|{}^\flat\theta\|\,{}^\flat R$ by using the conformal freedom to set ρ to unity, i.e., instead of (3.27) one chooses

$$\rho = 1, \tag{3.29}$$

which means that instead of (3.26) we shall have an explicit expression, namely,

$$\sigma^{2-n} = \|{}^\natural\theta\|, \tag{3.30}$$

by which the fiber measure density $\|{}^\natural\theta\|$ determines the conformal factor σ. With this choice, (3.23) simplifies, for $n\geqslant3$, to

$$\begin{aligned}\|\theta\|R + 2\|{}^\flat\theta\|\,{}^\flat\nabla_i\,{}^\flat\nabla^i\ln\sigma &= \|{}^\flat\theta\|\,{}^\flat R + \|{}^\flat\theta\|\{(2-n)({}^\flat\nabla_i\ln\sigma)\,{}^\flat\nabla^i\ln\sigma \\ &\quad+\tfrac14({}^\flat\nabla_i g_{rs})\,{}^\flat\nabla^i\,{}^\natural g^{rs} + \sigma^2 g_{rs}\omega_{ij}{}^r\omega^{ijs}\},\end{aligned}\tag{3.31}$$

while for (3.24) we obtain

$$\begin{aligned}\|\theta\|G_{XY}F_{\mu\nu}{}^X F^{\mu\nu Y} &= \|{}^\flat\theta\|G_{XY}\sigma^{-2}({}^\flat F_{ij}{}^X + 2\Phi_r{}^X\omega_{ij}{}^r) \\ &\quad\times({}^\flat F^{ijY} + 2\Phi_s{}^Y\omega^{ijs}) + 2\|{}^\flat\theta\|G_{XY} \\ &\quad\times{}^\natural g^{rs}({}^\flat\nabla_i\Phi_r{}^X)\,{}^\flat\nabla^i\Phi_s{}^Y,\end{aligned}\tag{3.32}$$

where we expanded the expressions for the $B_{ij}{}^X$ and the $E_{ir}{}^X$ in order to make it apparent that apart from a number of contributions of "harmonic" type, involving the base-space scalars g_{rs} and $\Phi_r{}^X$, we have recovered action contributions of the same form as we started with.

IV. THE REDUCED ACTION

The upshot of the preceding section is that, in the presence of the postulated Abelian invariance group, the variation problem set up in Sec. II can be replaced by an equivalent formulation in which the original set of m frame one-forms $\underline{\theta}^{\hat a}$ and q field one-forms $\underline{A}^X$ are replaced as independent variables by a (smaller) set of n ($=m-p$) base-space frame one-forms $\underline{\theta}^{\hat a}$ and a (larger) set of $(q+p)$ field one-forms consisting of the q base-space one-forms ${}^\flat\underline{A}^X$ together with the p base-space one-forms $\underline{\alpha}^r$. The $p(m+q)$ degrees of freedom lost in the reduction of the dimension of these (altogether $(m+q)$ one-forms from m to n are partly made up by the appearance of $\frac12 p(1+p+2q)$ additional variables, which we shall denote collectively by $\Phi^\blacklozenge$, introducing composite index variables, $\blacklozenge$, $\bullet$ running over $\frac12 p(1+p+2q)$ values representing and the pq components $\Phi_r{}^X$ and the $\frac12 p(p+1)$ distinct components g_{rs}, which have been demoted to the status of scalars (their contribution being transferred to the "matter" part of the Lagrangian) in the reduced formulation. [As an evocative illustration, in the case $p=2$, $q=5$, with the matrix g_{rs} having components $\left(\begin{smallmatrix}J & Q\\ Q & K\end{smallmatrix}\right)$ one could choose the first ten values of $\Phi^\blacklozenge$ by using the general purpose counting system that for $1\leqslant\blacklozenge\leqslant pq$ specifies $\Phi^{p(x-m)+r-n-1}=\Phi_r{}^x$, and then choose the remaining new scalar variables, i.e., the 11th, 12th, and 13th, to be $\Phi^{11}=J$, $\Phi^{12}=Q$, $\Phi^{13}=K$, this being itself an application of a general purpose counting system that, for $pq+1\leqslant\blacklozenge\leqslant pq+\frac12 p(p+1)$ and $s\geqslant r$ specifies $\Phi^{(1/2)(r-n-1)(2p+n-r)+s-n+pq}=g_{rs}$.] In this reformulated scheme the other lost geometric and "vectorial" degrees of freedom, accountable as $\frac12 p(m+p-1)$ degrees of fiber-frame rotation freedom, have been eliminated altogether. For the total Lagrangian density we shall have an equivalence relation (modulo additive divergence contributions) of the form

$$\mathscr{L}\cong{}^\flat\mathscr{L}, \tag{4.1}$$

where the new Lagrangian density ${}^\flat\mathscr{L}$ that is to be used in the reduced formulation will be expressible by a decomposition analogous to that of the original presentation (2.1) in the form

$${}^\flat\mathscr{L} = {}^\flat\mathscr{L}_{\text{geom}} + {}^\flat\mathscr{L}_{\text{vect}} + {}^\flat\mathscr{L}_{\text{matt}}, \tag{4.2}$$

in which the new vectorial and also, at least for $n\geqslant3$, the new geometric reduced contributions are of the same formal type as their original versions, while the remaining "material" contribution differs from its original version only by a term of purely harmonic type, meaning a contraction with respect to the base metric of a homogeneous quadratic in the gradients of the scalars, i.e.,

$${}^\flat\mathscr{L}_{\text{matt}} = \mathscr{L}_{\text{matt}} - \tfrac12\|{}^\flat\theta\|\mathscr{G}_{\blacklozenge\bullet}\,{}^\flat\nabla_i\Phi^\blacklozenge\,{}^\flat\nabla^i\Phi^\bullet, \tag{4.3}$$

where for $n\geqslant3$ the components of the composite matrix $\mathscr{G}_{\blacklozenge\bullet}$ may, if required, be read out from the explicit expansion

$$\begin{aligned}\mathscr{G}_{\blacklozenge\bullet}\,{}^\flat\nabla_i\Phi^\blacklozenge\,{}^\flat\nabla^i\Phi^\bullet &= \frac{G_{XY}}{4\pi G}\,{}^\natural g^{rs}({}^\flat\nabla_i\Phi_r{}^X)\,{}^\flat\nabla^i\Phi_s{}^Y \\ &\quad-\frac{1}{32\pi G}\left(\frac{{}^\natural g^{rs}\,{}^\natural g^{tu}}{n-2} + {}^\natural g^{rt}\,{}^\natural g^{su}\right) \\ &\quad\times({}^\flat\nabla_i g_{rs})\,{}^\flat\nabla^i g_{tu}.\end{aligned}\tag{4.4}$$

Thus (still subject to the proviso that $n\geqslant3$) the new (reduced) geometrical contribution will again take the standard Einstein–Hilbert form, i.e., we shall have

$${}^\flat\mathscr{L}_{\text{geom}} = (-\|{}^\flat\theta\|/16\pi G)\,{}^\flat R, \tag{4.5}$$

while the new (reduced) vectorial contribution, which is the principle subject of interest in the present discussion, will again (not only for $n\geqslant3$ but even for $n=2$) take a generalized Maxwellian form of the kind we started out with, i.e., we shall have

$${}^\flat\mathscr{L}_{\text{vect}} = (-\|{}^\flat\theta\|/16\pi G)\,{}^\flat G_{\Psi\Upsilon}\,{}^\flat F_{ij}{}^\Psi\,{}^\flat F^{ij\Upsilon}, \tag{4.6}$$

in which we have introduced late Greek capital indices Ψ, Υ with values ranging over the $p+q$ values $(n+1,\dots,m,m+1,\dots,m+q)$ covered jointly by the fiber-coordinate indices r, s and by the original field indices X, Y, and where definition of the fields ${}^\flat\underline{F}^\Psi$ and the corresponding potential covectors ${}^\flat\underline{A}^\Psi$ has been extrapolated naturally from the original range $(m+1,...,m+q)$ to the extended range $(n+1,...,m,m+1,...,m+q)$ of index values by defining the extra field variables to be

$$ {}^\flat A_i{}^r = \alpha_i{}^r, \quad {}^\flat F_{ij}{}^r = \omega_{ij}{}^r, \tag{4.7}$$

which by (3.17) [and of course (3.16)] is consistent with the basic exterior differentiation formula

$$ {}^\flat F_{ij}{}^\Psi = 2\,\partial_{[i}{}^\flat A_{j]}{}^\Psi. \tag{4.8}$$

For the general case, $n \geqslant 3$, the values of the extended $(p+q)\times(p+q)$ matrix of reduced coupling constants ${}^\flat G_{\Psi\Upsilon}$ may be read out explicitly from (3.31) as

$$\begin{aligned} &{}^\flat G_{XY} = \sigma^{-2} G_{XY}, \quad {}^\flat G_{Xr} = {}^\flat G_{rX} = 2\sigma^{-2} G_{XY}\Phi_r{}^Y, \\ &{}^\flat G_{rs} = \sigma^{-2}(g_{rs} + 4G_{XY}\Phi_r{}^X\Phi_s{}^Y). \end{aligned} \tag{4.9}$$

[In the special case $n=2$ the preceding formulas (4.9) would be modified only by the inclusion of an overall multiplicative factor ρ, but although (4.3) would be preserved, there would be a more substantial modification to (4.4) whose analog is obtainable from (3.28), while (4.5) would be subject to radical qualitative change since the Einstein–Hilbert contribution would be replaced simply by zero.]

In the reduced formulation that has just been constructed, reduced current and geometric source quantities may be naturally defined by

$$ {}^\flat \mathscr{J}_\Psi{}^i = \delta{}^\flat \mathscr{L}_{\text{matt}} / \delta{}^\flat A_i{}^\Psi, \tag{4.10}$$

and

$$ {}^\flat \mathscr{T}_{\hat a}{}^i = \delta{}^\flat \mathscr{L}_{\text{matt}} / \delta{}^\flat \theta_i{}^{\hat a}, \tag{4.11}$$

the latter being equivalent in the bosonic case to

$$ {}^\flat \mathscr{T}^{ij} = 2(\delta{}^\flat \mathscr{L}_{\text{matt}} / \delta{}^\flat g_{ij}). \tag{4.12}$$

These reduced current and geometric source can be evaluated explicitly in terms of the original (m-dimensional) current and geometric source quantities [as defined by (2.6) and (2.7) or (2.8)] as

$$ {}^\flat \mathscr{J}_r{}^i = 2(\mathscr{T}_r{}^i + \mathscr{J}_x{}^i\Phi_r{}^X), \quad {}^\flat \mathscr{J}_X{}^i = \mathscr{J}_X{}^i, \tag{4.13}$$

and

$$ {}^\flat \mathscr{T}^{ij} = \sigma^2 \mathscr{T}^{ij} + \|{}^\flat\theta\| \mathscr{G}_{\blacklozenge\bullet}({}^\flat\nabla^i\Phi^{\blacklozenge}){}^\flat\nabla^j\Phi^{\bullet}. \tag{4.14}$$

V. THE DIMENSIONAL AUGMENTATION (KALUZA–KLEIN) PROCEDURE

Having seen how the foregoing (Abelian) dimensional reduction scheme converts geometric (Einstein gravitational type) degrees of freedom to new (Maxwellian type) vectorial degrees of freedom, one is evidently able to proceed in the converse (Kaluza–Klein) sense whereby the vectorial degrees of freedom are converted to geometric degrees of freedom in a higher-dimensional bundle, ${}^\#\mathscr{M}$ say, whose total dimension will be $m+q$ where we recall that m was the dimension of the (physical) space that we started out with in Sec. II, and that q was the original number of Maxwellian-type vector fields $\underset{\sim}{A}{}^X$ which we now wish to absorb into the higher-dimensional geometry. This augmented manifold ${}^\#\mathscr{M}$ may be described (locally) in terms of new ignorable coordinates x^X for $X = m+1,...,m+q$ in addition to the original coordinates x^μ for $\mu = 1,...,m$ that were set up on $\mathscr{M}$ so that insofar as action on the relevant physical fields is concerned we shall have

$$ \partial_X = 0 \quad (X = m+1,...,m+q). \tag{5.1}$$

In order to set up the dimensionally augmented system, we shall need to introduce a new conformal factor ${}^\#\sigma$ defined analogously to our previous conformal factor σ. The appropriate metric on the artificially constructed $(m+q)$-dimensional bundle ${}^\#\mathscr{M}$ will have components which we shall denote by ${}^\# g_{\mathscr{I}\mathscr{J}}$, where we introduce script indices $\mathscr{I}$, $\mathscr{J}$ running over the whole range $1,...,m+q$, and which will evidently have to be given by a prescription of the form

$$\begin{aligned} &{}^\# g_{\mu\nu} = {}^\#\sigma^2(g_{\mu\nu} + 4G_{XY}A_\mu{}^X A_\nu{}^Y), \\ &{}^\# g_{\mu X} = 2{}^\#\sigma^2 G_{XY}A_\mu{}^Y, \quad {}^\# g_{XY} = {}^\#\sigma^2 G_{XY}. \end{aligned} \tag{5.2}$$

The applicability of this Kaluza–Klein ansatz does not of course depend on the existence of the invariance group and the associated dimensional reduction that was described in Secs. III and IV, but when the latter is present the prescription (5.2) can be expressed in terms of the corresponding reduced quantities in the equivalent alternative form

$$\begin{aligned} &{}^\# g_{ij} = {}^\#\sigma^2\sigma^2({}^\flat g_{ij} + 4{}^\flat G_{\Psi\Upsilon}\,{}^\flat A_i{}^\Psi\,{}^\flat A_j{}^\Upsilon), \\ &{}^\# g_{i\Psi} = 2{}^\#\sigma^2\sigma^2\,{}^\flat G_{\Psi\Upsilon}\,{}^\flat A_i{}^\Upsilon, \quad {}^\# g_{\Psi\Upsilon} = {}^\#\sigma^2\sigma^2\,{}^\flat G_{\Psi\Upsilon}. \end{aligned} \tag{5.3}$$

The appropriate value for the conformal factor will be given in terms of the determinant $|G|$ of the $q\times q$ coupling matrix G_{XY} by the relation

$$ {}^\#\sigma^{-2} = \|G\|^{1/(d-2)}, \tag{5.4}$$

where

$$ d = m+q = n+p+q \tag{5.5}$$

is the total dimension of the augmented bundle space.

For the Lagrangian density we shall then have an equivalence relation (modulo a divergence) of the form

$$ \mathscr{L} \cong {}^\#\mathscr{L}, \tag{5.6}$$

where the higher $(m+q)$-dimensional version ${}^\#\mathscr{L}$ is given simply by

$$ {}^\#\mathscr{L} = {}^\#\mathscr{L}_{\text{geom}} + {}^\#\mathscr{L}_{\text{matt}}, \tag{5.7}$$

the vector part having disappeared, where the geometric part has the usual Einstein–Hilbert form,

$$ {}^\#\mathscr{L}_{\text{geom}} = \|{}^\#\theta\|{}^\# R \tag{5.8}$$

in terms of the $(m+q)$-dimensional curvature scalar, ${}^\# R$ and the $(m+q)$-dimensional measure density, $\|{}^\#\theta\| = \|{}^\# g\|^{1/2}$.

There is, however, a feature of this construction which is for many purposes undesirable, which is that for the matter contribution ${}^\#\mathscr{L}_{\text{matt}}$ there will be a d-dimensionally noncovariant adjustment term of the form

$$ {}^\#\mathscr{L}_{\text{matt}} - \mathscr{L}_{\text{matt}} = \tfrac{1}{2}\|{}^\#\theta\|\mathscr{G}_{\circ\diamond}({}^\#\nabla_{\mathscr{I}}\,{}^\#\Phi^{\circ}){}^\#\nabla^{\mathscr{I}}\,{}^\#\Phi^{\diamond} \tag{5.9}$$

in which there are $\frac{1}{2}q(q+1)$ new field quantities, which have been denoted collectively by ${}^\#\Phi^\circ$ which from an $(m+q)$-dimensional point of view are not scalars but tensor components with respect to a privileged fibration, namely the independent metric components ${}^{\#}g_{XY}$. Its explicit form is

$$\mathscr{G}_{\bigcirc\Diamond}({}^{\#}\nabla_{\mathscr{I}}\ {}^{\#}\Phi^{\bigcirc}){}^{\#}\nabla^{\mathscr{I}}\ {}^{\#}\Phi^{\Diamond}=\left\{\frac{{}^{\natural\#}g^{XY}\ {}^{\natural\#}g^{WZ}}{m-2}+{}^{\natural\#}g^{XW}\ {}^{\natural\#}g^{YZ}\right\}\times\frac{({}^{\#}\nabla_{\mathscr{I}}\ {}^{\#}g_{XY}){}^{\#}\nabla^{\mathscr{I}}\ {}^{\#}g_{WZ}}{16\pi G},\qquad(5.10)$$

where, in accordance with the convention introduced in Sec. III, the prefixed "natural" symbol indicates that the ${}^{\natural\#}g^{XY}$ are the components of the natural inverse of the submatrix ${}^{\#}g_{XY}$ as distinct from the submatrix components ${}^{\#}g^{XY}$ of the full matrix inverse ${}^{\#}g^{\mathscr{I}\mathscr{I}}$ of ${}^{\#}g_{\mathscr{I}\mathscr{I}}$. When the Kaluza–Klein ansatz is used as a guide to the construction of new theories (rather than as a mathematically useful reformulation of existing theories, which is the point of view corresponding to the spirit of the present work) one normally imposes the requirement that ${}^{\#}\mathscr{L}_{\rm matt}$ be d-dimensionally covariant, which means that the offending contribution on the right of (5.9) must be canceled out from ${}^{\#}\mathscr{L}_{\rm matt}$ by a corresponding term in $\mathscr{L}_{\rm matt}$. In this case the extra contribution will turn up in the form

$$\mathscr{L}_{\rm matt}-{}^{\#}\mathscr{L}_{\rm matt}=-(\|\theta\|/2)\mathscr{G}_{\bigcirc\Diamond}(\nabla_{\mu}\ {}^{\#}\Phi^{\bigcirc})\nabla^{\mu}\ {}^{\#}\Phi^{\Diamond}\qquad(5.11)$$

as a well behaved scalar contribution in the lower-dimensional formulation, being given explicitly in satisfactorily m-dimensionally covariant form by

$$\mathscr{G}_{\bigcirc\Diamond}(\nabla_{\mu}\ {}^{\#}\Phi^{\bigcirc})\nabla^{\mu}\ {}^{\#}\Phi^{\Diamond}=\left\{\frac{3(m-2)G^{XY}G^{WZ}}{4(m-1)^2}+G^{XW}G^{YZ}\right\}\times\frac{(\nabla_{\mu}G_{XY})\nabla^{\mu}G_{WZ}}{16\pi G},\qquad(5.12)$$

where the G^{XY} are the components of the inverse of the coupling matrix G_{XY}. The contribution (5.11) will thus be interpretable as representing an effective action for the coupling parameters as dynamic scalar fields. However, in the alternative approach exemplified by standard Einstein–Maxwell theory, in which one wishes to have the G_{XY} as fixed constants, it is the apparent dynamic contribution on the right-hand side of (5.11) that must be supposed to be canceled by a corresponding counterterm in ${}^{\#}\mathscr{L}_{\rm matt}$, whose covariance in d dimensions will then be destroyed by the effective presence of the term given on the right-hand side of (5.9).

Insomuch as the present general investigation is essentially concerned not with the scalar contributions but rather with the vectorial contributions, the results to be obtained will be valid for both the Kaluza–Klein type and the "pure" (scalar-free) Einstein–Maxwell type theories, since the effect of the distinction on the source contributions is trivial. More specifically, the inclusion of the extra Kaluza–Klein type scalar field contribution (5.12) will in no way alter the sources (4.13) for the base-space projected vector fields. To see this one may start by remarking that it is immediately obvious that the primary source current is quite unaffected by the change from $\mathscr{L}$ to ${}^{\#}\mathscr{L}$, i.e., one has

$$\delta^{\#}\mathscr{L}_{\rm matt}/\delta A_{\mu}{}^{X}=\mathscr{J}_{X}{}^{\mu}\qquad(5.13)$$

with the current density $\mathscr{J}_{X}{}^{\mu}$ as before, while for the energy momentum we shall have

$$2(\delta^{\#}\mathscr{L}_{\rm matt}/\delta g^{\mu\rho})={}^{\#}\mathscr{T}^{\mu\rho},\qquad(5.14)$$

where the new tensor density ${}^{\#}\mathscr{T}^{\mu\rho}$, as so defined, differs from the original material energy contribution $\mathscr{T}^{\mu\rho}$ only by a term whose mixed (covariant and contravariant) coordinate version, which is the one that concerns us directly as a source contribution, will be expressible directly in the form

$$\mathscr{T}_{\rho}{}^{\mu}-{}^{\#}\mathscr{T}_{\rho}{}^{\mu}=\|\theta\|\mathscr{G}_{\bigcirc\Diamond}(\nabla^{\mu}\ {}^{\#}\Phi^{\bigcirc})\partial_{\rho}\ {}^{\#}\Phi^{\Diamond}.\qquad(5.15)$$

It is therefore immediately apparent that when the covariant index ρ lies in the ignorable range ($r=n+1,...,m$) the manifest stationarity condition (3.1) as applied to the scalars ${}^{\#}\Phi^{\Diamond}$ (i.e., the G_{XY}) implies that the distinction will disappear, i.e., for these coordinate values we shall have

$${}^{\#}\mathscr{T}_{r}{}^{\mu}=\mathscr{T}_{r}{}^{\mu},\qquad(5.16)$$

which effectively completes the demonstration of the point we wish to make, namely that the effective presence or absence of a Kaluza–Klein type scalar contribution of the form (5.11) will not only be irrelevant for the primary source currents, as defined by (2.6) and given by (5.13), but that it will also be irrelevant for the reduced source currents as defined by (4.10) and given by (4.13)

VI. THE COMOVEMENT CONDITION AND THE DUAL TRIVECTOR FORMULATION

After these preliminary generalities, we now concentrate our attention on the particular class of situations with which we shall be concerned throughout the remainder of the present work, namely those for which the sources included in $\mathscr{L}_{\rm matt}$ are effectively *comoving* with the generators of the symmetry group, in the sense that the surfaces of transitivity of the group contain the directions of the current densities $\mathscr{J}_{X}{}^{\mu}$ and that they should be spanned by a subset of eigenvectors of the energy momentum density $\mathscr{T}^{\mu\nu}$, which means that only components of the form $\mathscr{J}_{X}{}^{r}$, $\mathscr{T}_{s}{}^{r}$, $\mathscr{T}_{j}{}^{i}$ $(i,j\leqslant n<r,s\leqslant n+p=m<X\leqslant m+q)$ should be present, i.e.,

$$\mathscr{J}_{X}{}^{i}=0,\qquad(6.1)$$

$$\mathscr{T}_{r}{}^{i}=0.\qquad(6.2)$$

This *comovement condition* is equivalent by (4.13) to the condition that the effective base-space currents ${}^{\flat}\mathscr{J}_{X}{}^{i}$ and ${}^{\flat}\mathscr{J}_{r}{}^{i}$ should vanish, i.e., in the condensed notation scheme given by (4.10),

$${}^{\flat}\mathscr{J}_{\Xi}{}^{i}=0\qquad(6.3)$$

for the whole range $n<\Xi\leqslant m+q$. Under these conditions it can be seen that the dynamic field equations for the unknowns ${}^{\flat}A_{i}{}^{\Xi}$ will be obtainable directly from the action contribution (4.6) in the form

$${}^{\flat}\nabla_{j}\ {}^{\flat}G_{\Xi\Upsilon}\ {}^{\flat}F^{ij\Upsilon}=0,\qquad(6.4)$$

which may be written out more explicitly as the distinct sets of equations

$$ {}^\flat\nabla_j \sigma^{-2}(g_{rs}\omega^{ijr} + 2\Phi_r{}^X G_{XY} B^{ijY}) = 0, \tag{6.5} $$

$$ {}^\flat\nabla_j \sigma^{-2} G_{XY} B^{ijY} = 0, \tag{6.6} $$

for the $\alpha_i{}^r$ and the ${}^\flat A_i{}^X$, respectively.

The generalized Maxwell equations (6.4) may as usual be interpreted as Poincaré type integrability conditions: when they are satisfied, i.e., when the comovement requirement (6.3) holds, then the fields ${}^\flat F_{ij}{}^\Xi$ will not only be derivable from the one-form potentials $A_i{}^\Xi$ from which they were originally constructed, but, except in the $n = 2$ case which will be left aside for special treatment later on, they will also be derivable from a set of base-space *trivectors* (antisymmetric third-order contravariant tensors) with the components $\mathscr{W}_\Xi{}^{ijk}$ via a set of generalized divergence relations of the form

$$ {}^\flat G_{\Xi\Upsilon}\, {}^\flat F^{ij\Upsilon} = {}^\flat\nabla_k \mathscr{W}_\Xi{}^{ijk}, \tag{6.7} $$

or equivalently in the more explicit form

$$ \sigma^{-2}(g_{rs}\omega^{ijs} + 2\Phi_r{}^X G_{XY} B^{ijY}) = {}^\flat\nabla_k \mathscr{W}_r{}^{ijk}, \tag{6.8} $$

$$ \sigma^{-2} G^{-1} G_{XY} B^{ijY} = {}^\flat\nabla_k\, \Psi_X{}^{ijk}, \tag{6.9} $$

where the full set of trivectors $\mathscr{W}_\Xi{}^{ijk}$ is considered as a combination of the strictly geometric subset $\mathscr{W}_r{}^{ijk}$ as defined by (6.8) and the subset of magnetic potential trivectors ${}^\flat\Psi_X{}^{ijk}$ as defined by (6.9), the specification of the full set in the condensed expression (6.7) being completed by setting

$$ \mathscr{W}_X{}^{ijk} = G\, {}^\flat\Psi_X{}^{ijk}. \tag{6.10} $$

The preceding relations may be solved in the form

$$ {}^\flat F^{ij\Upsilon} = {}^\flat G^{\Xi\Upsilon}\, {}^\flat\nabla_k \mathscr{W}_\Upsilon{}^{ijk}, \tag{6.11} $$

with the usual convention that the ${}^\flat G^{\Xi\Upsilon}$ denote the components of the matrix inverse to the ${}^\flat G_{\Xi\Upsilon}$, which gives the explicit expressions

$$ \omega^{ijr} = \sigma^2\, {}^\natural g^{rs}({}^\flat\nabla_k \mathscr{W}_s{}^{ijk} - 2G\Phi_s{}^X\, {}^\flat\nabla_k\, {}^\flat\Psi_X{}^{ijk}), \tag{6.12} $$

$$ \begin{aligned} {}^\flat F^{ijX} = \sigma^2\{ & G(G^{XY} + 4\, {}^\natural g^{rs}\Phi_r{}^X\Phi_s{}^Y)\, {}^\flat\nabla_k\, {}^\flat\Psi_Y{}^{ijk} \\ & - 2\Phi_r{}^X\, {}^\natural g^{rs}\, {}^\flat\nabla_k \mathscr{W}_s{}^{ijk}\}. \end{aligned} \tag{6.13} $$

Under the circumstances that have just been described we may carry out a transformation that generalizes the one of the type introduced by Geroch[20] in the $n = 3$ case (which was itself an analog of the transformation introduced by Ernst[21] in the exceptional $n = 2$ case that we have had to leave aside for the time being) so as to replace the original formulation of the system by an equivalent *dual reformulation* in which the original independent one-form variables ${}^\flat A_i{}^\Xi$ (i.e., the $\alpha_i{}^r$ and the ${}^\flat A_i{}^X$) are replaced by the dual three-vector variables $\mathscr{W}_\Xi{}^{ijk}$ (i.e., the $\mathscr{W}_r{}^{ijk}$ and the ${}^\flat\Psi_X{}^{ijk}$), which can be done for any base-space dimension $n > 2$, but which is of course particularly advantageous in the $n = 3$ case for which the three-vectors will be algebraically dual to scalars, so that the reformulation will effectively reduce the original (linear) vector system to a scalar system of (linear) harmonic type.

The duality transformation of the system can be carried through directly at the level of the Lagrangian (4.6) [without explicitly invoking the field equations (6.4)] by proceeding in two steps (in the manner used by Breitenlohner, Maison, and Gibbons[10] in the three-dimensional case) of which the first consists of replacing the original potentials ${}^\flat A_i{}^\Xi$ by the corresponding fields ${}^\flat F_{ij}{}^\Xi$ as independent variables, using corresponding Lagrange multipliers to enforce the required integrability conditions

$$ \partial_{[i}\, {}^\flat F_{jk\,]}{}^\Xi = 0 \tag{6.14} $$

for the existence of the former. The possibility of doing this is of course entirely dependent on the *comovement postulate* (6.3) which by (4.10) ensures that the ${}^\flat A_i{}^\Xi$ do not appear anywhere in ${}^\flat\mathscr{L}$ except in the combination ${}^\flat F_{ij}{}^\Xi$. The equivalent modified Lagrangian contribution

$$ {}^\flat\mathscr{L}'_{\text{vect}} \cong {}^\flat\mathscr{L}_{\text{vect}} - (\|{}^\flat\theta\|/8\pi\mathrm{G})\, \mathscr{W}_\Xi{}^{ijk}\partial_{[i}\, {}^\flat F_{jk\,]}{}^\Xi \tag{6.15} $$

will automatically give equations of the form

$$ {}^\flat\nabla_k \mathscr{W}_\Xi{}^{ijk} = -8\pi G(\delta\, {}^\flat\mathscr{L}_{\text{vect}}/\delta\, {}^\flat F_{ij}{}^\Xi) \tag{6.16} $$

from which it can be seen that the Lagrange multipliers $\mathscr{W}_\Xi{}^{ijk}$ may be identified with the trivector potentials introduced in (6.7). It is to be remarked that in order to satisfy the requirements of the variational principle it is not sufficient merely to ensure that the fields ${}^\flat F_{ij}{}^\Xi$ are compatible just with the local existence of the ${}^\flat A_i{}^\Xi$ which is all that the basic Poincaré lemma provides from (6.14): it is necessary to ensure that for any variation $\delta\, {}^\flat F_{ij}{}^\Xi$ with support confined to a small neighborhood it will be possible to choose a gauge such that the corresponding variation $\delta\, {}^\flat A_i{}^\Xi$ also has support confined to a small surrounding neighborhood. The gauge adjustment that may be needed to ensure that $\delta\, {}^\flat A_i{}^\Xi$ vanishes in an outer shell surrounding the original neighborhood will always be possible if the shell is simply connected, and for $n \geqslant 3$ the well known simple connectivity property of spheres takes care of this. However, this would not work in the case $n = 2$, which has already been provisionally excluded, since in this case the surrounding shell becomes a ring, so that the required locally supported variation $\delta\, {}^\flat A_i{}^\Xi$ will not in general exist, even though in this $n = 2$ case the local integrability condition (6.14) will always hold as a trivial identity.

To complete the duality formulation as described, which for the foregoing reason is possible only for base-space dimension $n \geqslant 3$, one now reduces the number of degrees of freedom (which have temporarily been augmented by the introduction of the $\mathscr{W}_\Xi{}^{ijk}$) by restraining the variables to satisfy the field equations (6.16) in advance of the variation, which means that the ${}^\flat F_{ij}{}^\Xi$ revert to their original status of secondary (derived) quantities, being now considered as defined by the solution (6.11) of (6.16). It is convenient at this stage also to fix the additive divergence ambiguity in the equivalence relation (6.15) so as to obtain a Lagrangian density contribution involving only first-order derivatives of the independent fields, which gives

$$ {}^\flat\mathscr{L}' = {}^\flat\mathscr{L} + (\|{}^\flat\theta\|/8\pi G)\, {}^\flat F_{jk}{}^\Xi\, {}^\flat\nabla_i \mathscr{W}_\Xi{}^{ijk}, \tag{6.17} $$

where the ${}^\flat F_{ij}{}^\Xi$ are now considered to be defined by (6.11). Algebraically the effect of this is simply to reverse the sign of the vectorial contribution: one obtains a total Lagrangian ${}^\flat\mathscr{L}'$ in the form

$$ {}^\flat\mathscr{L}' = {}^\flat\mathscr{L}_{\rm geom} + {}^\flat\mathscr{L}'_{\rm vect} + {}^\flat\mathscr{L}_{\rm matt}, \qquad (6.18) $$

which differs from (4.2) only in that we have

$$ {}^\flat\mathscr{L}'_{\rm vect} = -{}^\flat\mathscr{L}_{\rm vect}. \qquad (6.19) $$

For variational purposes the explicit form of the new vectorial contribution in the dual reformulation will be

$$ {}^\flat\mathscr{L}_{\rm vect} = (\|{}^\flat\theta\|/16\pi G)\, {}^\flat G^{\Xi\Upsilon}({}^\flat\nabla_k \mathscr{W}_{\Xi}{}^{ijk})\, {}^\flat\nabla_h \mathscr{W}_{\Upsilon ij}{}^h \qquad (6.20) $$

or in more detail

$$ \begin{aligned} {}^\flat\mathscr{L}_{\rm vect} = (\|{}^\flat\theta\|\sigma^2/16\pi)\{&GG^{XY}({}^\flat\nabla_k\, {}^\flat\Psi_X{}^{ijk})\, {}^\flat\nabla_h\, {}^\flat\Psi_{Yij}{}^h \\ &+ G^{-1}\, {}^\natural g^{rs}({}^\flat\nabla_k \mathscr{W}_r{}^{ijk} - 2G\Phi_r{}^X\, {}^\flat\nabla_k\, {}^\flat\Psi_X{}^{ijk}) \\ &\times({}^\flat\nabla_h \mathscr{W}_{rij}{}^h - 2G\Phi_r{}^X\, {}^\flat\Psi_{Xij}{}^h)\}. \end{aligned} \qquad (6.21) $$

While, as we have seen, it can be done for any base space dimension $n \geqslant 2$, this dual reformulation is most obviously advantageous in the lowest allowed base dimension, namely $n = 3$, for which we may simply write

$$ \mathscr{W}_r{}^{ijk} = {}^*\mathscr{W}_r\, {}^\flat\epsilon^{ijk}, \quad {}^\flat\Psi_X{}^{ijk} = {}^*\Psi_X\, {}^\flat\epsilon^{ijk}, \qquad (6.22) $$

where ${}^\flat\epsilon^{ijk}$ is the three-dimensional alternating tensor determined by the three-dimensional base metric ${}^\flat g_{ij}$, and where the ${}^*\mathscr{W}_r$ and the ${}^*\Psi_X$ are the simple base-space scalars, so that the preceding expression can then be reduced to the (linear) purely harmonic scalar form

$$ \begin{aligned} {}^\flat\mathscr{L}'_{\rm vect} = (\|{}^\flat\theta\|\sigma^2/8\pi)\{&GG^{XY}({}^\flat\nabla_k\, {}^*\Psi_X)\, {}^\flat\nabla^k\, {}^*\Psi_Y \\ &+ G^{-1}\, {}^\natural g^{rs}({}^\flat\nabla_k\, {}^*\mathscr{W}_r - 2\Phi_r{}^X\, {}^\flat\nabla_k\, {}^*\Psi_X) \\ &\times({}^\flat\nabla^k\, {}^*\mathscr{W}_s - 2\Phi_s{}^Y\, {}^\flat\nabla^k\, {}^*\Psi_Y)\}. \end{aligned} \qquad (6.23) $$

Although the $n = 2$ case that we have left aside until now does not have a differentially dual reformulation of the kind just described, it can also be simplified by a reformulation in terms of a set of scalars, namely the quantities ${}^*\omega^r$ and ${}^{*\flat}F^X$, or collectively in condensed notation ${}^*\omega^\Xi$, as defined by the purely algebraic duality transformations

$$ {}^*\omega^\Xi = \tfrac{1}{2}\, {}^\flat\epsilon^{ij}\, {}^\flat F_{ij}{}^\Xi = {}^\flat\nabla_i\, {}^*\alpha^{i\Xi}, \qquad (6.24) $$

where the two-dimensional contravariant vector potentials are defined by

$$ {}^*\alpha^{i\Xi} = {}^\flat\epsilon^{ij}\, {}^\flat A_j{}^\Xi. \qquad (6.25) $$

Thus in this special (two-dimensional base) case one obtains

$$ \mathscr{L}_{\rm vect} = (-\|{}^\flat\theta\|/8\pi G)\, {}^\flat G_{\Xi\Upsilon}\, {}^*\omega^\Xi\, {}^*\omega^\Upsilon, \qquad (6.26) $$

which implies

$$ \delta\, {}^\flat\mathscr{L}_{\rm vect}/\delta\, {}^*\alpha^{i\Xi} = (\|{}^\flat\theta\|/4\pi G)\, {}^\flat\nabla_i\, {}^{*\flat}G_{\Xi\Upsilon}\, {}^*\omega^\Upsilon \qquad (6.27) $$

so the corresponding field equations will be expressible on first integrated form as

$$ {}^*\omega^\Xi = {}^\flat G^{\Xi\Upsilon} C_\Upsilon, \qquad (6.28) $$

where the C_Υ are a set of arbitrary constants of integration.

VII. GLOBAL CONDITIONS FOR ORTHOGONAL TRANSITIVITY

The remarks at the end of the preceding section lead in naturally to the main purpose of the present work, which is the investigation of the circumstances under which the group action under consideration (as introduced in Sec. III) may be expected to be *orthogonally transitive* in the sense[4] that there exist families of (complementary, i.e., n-dimensional) surfaces orthogonal to the surfaces of transitivity over which the group acts, both in the original [$(m = n + p)$-dimensional] space with metric $g_{\mu\rho}$, and in a stronger sense of the term, in the [$(d = m + q)$-dimensional] augmented bundle space with metric ${}^\sharp g_{IJ}$. Orthogonal transitivity, which we may refer to unambiguously in the present context just as "orthogonality," is thus equivalent to removability of the quantities $\alpha_i{}^r$ (by coordinate transformations) and also, when interpreted in the stronger sense, removability of the quantities ${}^\flat A_i{}^X$ (by gauge transformations), for which necessary conditions, which are also locally—and in a simply connected region also globally—sufficient are the requirements that the corresponding exterior derivatives should vanish, so that for orthogonality in the weak (purely geometric) sense we should have

$$ \omega_{ij}{}^r = 0, \qquad (7.1) $$

while for orthogonality in the strong sense we should also have

$$ {}^\flat F_{ij}{}^X = 0, \qquad (7.2) $$

which by (3.15) is equivalent subject to (7.1) to the vanishing of the generalized magnetic type field contribution $B_{ij}{}^X$.

Since the field equations for the $\alpha_i{}^r$ and the ${}^\flat A_i{}^X$ are obtainable from (4.6) and (4.10) in the combined form

$$ (\|{}^\flat\theta\|/4\pi G)\, {}^\flat\nabla_j\, {}^\flat G_{\Xi\Upsilon}\, {}^\flat F^{ij\Upsilon} = {}^\flat\mathscr{J}_\Xi{}^i, \qquad (7.3) $$

one sees that the *comoving source conditions* (6.3), i.e.,

$$ {}^\flat\mathscr{J}_r{}^i = 0, \qquad (7.4) $$

$$ {}^\flat\mathscr{J}_X{}^i = 0, \qquad (7.5) $$

are automatic consequences of the respective orthogonality conditions (7.1) and (7.2), whose combination is given in our condensed notation scheme as

$$ {}^\flat F_{ij}{}^\Xi = 0. \qquad (7.6) $$

Our present purpose is to investigate the circumstances under which these comovement conditions, (7.4) and (7.5) in combination, as expressed together in condensed notation by the comovement condition (6.3) postulated at the outset of the preceding section, should not merely be locally necessary but also globally sufficient to guarantee the strong orthogonality property (7.6).

Before moving to the general case with base dimension $n \geqslant 3$, we remark that in the special case with base dimension $n = 2$ it can be seen directly from the work at the end of the previous section that (as has been known since the early general study by the present author,[4] following the demonstration of the original orthogonal transitivity theorem by Papapetrou[3] for the special case of a pure Einstein vacuum in a four-dimensional stationary axisymmetric space-time) only very weak boundary conditions will be needed for (6.3) [i.e., the combination of (7.4) and (7.5)] to ensure the orthogonality property (7.6) [i.e., the combination of (7.1) and

(7.2)] since all that will be needed is that the boundary conditions at some single point should be sufficiently restrictive to ensure that the constants of integration C_Υ introduced in (6.28) should vanish. In the typical application to stationary-axisymmetric astrophysical models, including rotating stellar and black-hole equilibrium states, the connected region under investigation will include (and therefore the two-dimensional quotient space will be partially bounded by) an axis of rotation symmetry on which[4] the relevant fields, $\omega_{ij}{}^r$ and ${}^\flat F_{ij}{}^X$ in the present notation scheme, must necessarily vanish identically, which is sufficient to ensure the vanishing of the relevant constants C_r and C_X and hence that the required result (7.6) should hold over the entire (connected) region provided the comovement condition (6.3) (which in this context is interpretable[7] as a source flux circularity condition) is known to hold throughout.

For the general case (as first systematically investigated here) with base-space dimension $n \geqslant 3$, the circumstances under which (6.3) [i.e., (7.4) and (7.5)] should be sufficient to ensure (7.6) [i.e., (7.1) and (7.2)] are much less general than for $n = 2$, but they are nevertheless sufficiently extensive to be of widespread practical interest. The prototype result (the analog of Papapetrou's theorem for the $n = 2$, $m = 4$ case) is a well known theorem of Lichnerowicz[6] in the purely gravitational $n = 3$, $m = 4$ case for timelike one-dimensional isometry group trajectories, our general comovement condition being interpretable in such a context as a *staticity condition*. The original Lichnerowicz theorem required fairly severe boundary conditions involving asymptotic flatness at spatial infinity, and also Euclidean spatial (quotient space) topology, but it was later remarked by Hawking[5] that the latter requirement could be relaxed to allow for the presence of a central black hole subject to a further staticity restriction on the horizon. The more complicated situation arising in the presence of electromagnetic effects has been considered on two occasions in summer school proceedings by the present author[7,8] (a nontrivial sign error in the earlier one having been corrected in the latter). The present work shows how these $n = 3$ examples can be considered as special applications of general results valid for any base-space dimension $n \geqslant 3$.

The general method on which this work is based is the construction of linear combinations of the field equations in such a way as to express a positive or negative definite function of the field variables $\omega_{ij}{}^r$ and ${}^\flat F_{ij}{}^X$ in terms of a divergence whose space integral converts to a surface contribution that will be eliminated by the imposition of suitable boundary conditions. In order to have as much flexibility as possible in seeking an appropriate form for the surface contribution, the alternative dual reformulation described in the previous section turns out to be particularly useful. Starting from the full set of generalized Maxwell type equations (6.4), (6.14) that are obtained from the ordinary or dual vectorial contributions [(4.6) or (6.20)] to the Lagrangian, the most obviously promising divergence combination, Σ say, obtained by contracting the full set of field equations with the corresponding (covector or trivector) potentials is proportional to the Lagrangian contribution itself, being given by the homogeneous quadratic expressions

$$\begin{aligned}\Sigma &= {}^\flat F_{ij}{}^{\Xi}\, {}^\flat G_{\Xi\Upsilon}\, {}^\flat F^{ij\,\Upsilon} \\ &= ({}^\flat\nabla_k \mathscr{W}_{\Xi\ ij}{}^k)\, {}^\flat G^{\Xi\Upsilon}\, {}^\flat\nabla_h \mathscr{W}_\Upsilon{}^{ijk} \\ &= \sigma^{-2}\{ g_{rs}\omega_{ij}{}^r \omega^{ij\,s} + G_{XY} B_{ij}{}^X B^{ij\,Y}\},\end{aligned} \tag{7.7}$$

which (by the field equations) can be converted into a divergence in two essentially different ways expressible by

$$\Sigma = 2\, {}^\flat\nabla_i ({}^\flat A_j{}^{\Xi}\, {}^\flat G_{\Xi\Upsilon}\, {}^\flat F^{ij\,\Upsilon}) = {}^\flat\nabla_k ({}^\flat F_{ij}{}^{\Xi} \mathscr{W}_{\Xi}{}^{ijk}). \tag{7.8}$$

It can be seen that the quadratic expressions (7.7) will have a positivity or negativity property of the required type provided we have separate positivity or negativity properties for each of the matrices ${}^\flat g_{ij}$ (the $n \times n$ base metric) and ${}^\flat G_{\Xi\Upsilon}$ [the $(p+q) \times (p+q)$ reduced coupling matrix] the latter requirement being equivalent [by the defining relation (4.9)] to demanding the *same* positivity or negativity property for *both* the physical coupling matrix G_{XY} and the fiber (ignorable coordinate) component metric submatrix g_{rs}. It is clear, however, that in an ordinary space-time situation, with an indefinite-signature metric $g_{\mu\rho}$, when the base metric ${}^\flat g_{ij}$ is positive definite then the fiber metric g_{rs} will be able to have a definite signature only if it is negative (which in the normal Lorentzian case requires $p = 1$, implying that it will only have a single component, g_{nn} say) so that the above requirements could only be achieved for a coupling matrix G_{XY} of *negative* type. In a physical situation of the usual kind, for which G_{XY} would be of *positive* type, the divergence relations (7.8) would therefore be unable to provide any orthogonal transitivity theorems in ordinary space-time, though they could be applied successfully to the situation obtained by performing a Wick type (complex) rotation so as to make $g_{\mu\rho}$ positive definite (Euclidian signature): thus we get a class of orthogonal transitivity theorems for caloron-type situations with periodic boundary condition in the direction of what (before the Wick rotation) had been the time direction, subject to suitable asymptotic flatness conditions in the other spatial directions.

Returning our attention to the normal physical situation in an indefinite signature space-time with a positive coupling matrix G_{XY} (as defined in terms of the convention that we use a positive signature for space as distinguished from time), we see that a rather more elaborate alternative to (7.7) will be needed to get any orthogonal transitivity theorem of the kind we are seeking. However the last (most detailed) expression on the right-hand side of (7.7) provides the clue to a natural way of constructing an alternative divergence relation more suitable for this purpose. Instead of just contracting all the ${}^\flat F_{ij}{}^{\Xi}$ with the divergences of the twist potentials $\mathscr{W}_{\Xi}{}^{ijk}$ using the same sign throughout, we switch the relative sign of the subset of purely geometric origin, thereby obtaining a new homogeneous quadratic function $\widetilde{\Sigma}$ of the fields that is given by

$$\begin{aligned}\widetilde{\Sigma} &= {}^\flat F_{ij}{}^X\, {}^\flat G_{XY}\, {}^\flat F^{ij\,Y} - \omega_{ij}{}^r\, {}^\flat G_{rs} \omega^{ij\,s} \\ &= G^2({}^\flat\nabla_k\, {}^\flat\Psi_{X\ ij}{}^k)\, {}^\flat G^{XY}\, {}^\flat\nabla_h\, {}^\flat\Psi_Y{}^{ijk} \\ &\quad - ({}^\flat\nabla_k \mathscr{W}_{r\ ij}{}^k)\, {}^\flat G^{rs}\, {}^\flat\nabla_h \mathscr{W}_s{}^{ijk}\end{aligned} \tag{7.9}$$

and which can also be converted (using the field equations) into a divergence in two essentially different ways, which are expressible by

$$\widetilde{\Sigma} = 2\, {}^{\flat}\nabla_i \{ ({}^{\flat}A_j{}^X\, {}^{\flat}G_{X\Xi} - \alpha_j{}^r\, {}^{\flat}G_{r\Xi})\, {}^{\flat}F^{ij\,\Xi} \}$$

$$= {}^{\flat}\nabla_k (G\, {}^{\flat}F_{ij}{}^X\, {}^{\flat}\Psi_X{}^{ijk} - \omega_{ij}{}^r \mathscr{W}_r{}^{ijk}). \qquad (7.10)$$

We now have a quantity whose expression (7.9) can be seen to have a positivity or negativity property of the required kind for a positive definite base metric ${}^{\flat}g_{ij}$ provided the submatrices ${}^{\flat}G_{XY}$ and ${}^{\flat}G_{rs}$ (or equivalently ${}^{\flat}G^{XY}$ and ${}^{\flat}G^{rs}$) have the *opposite* positive or negative definiteness properties. Thus in the usual case where the coupling matrix G_{XY}—and hence also the reduced coupling submatrix ${}^{\flat}G_{XY}$—is positive definite, we need that the submatrix ${}^{\flat}G_{rs}$ should be negative definite (these properties being equivalent to the condition that the submatrices ${}^{\flat}G^{XY}$ and ${}^{\flat}G^{rs}$ should be, respectively, positive and negative definite). More explicitly this means that to obtain a positive definite right-hand side for (7.9) subject to the usual requirement that the base metric ${}^{\flat}g_{ij}$ be itself positive definite with a positive definite coupling matrix G_{XY} it is necessary and sufficient to satisfy the negativity condition

$$(g_{rs} + 4G_{XY}\Phi_r{}^X\Phi_s{}^Y)\zeta^r\zeta^s < 0 \qquad (7.11)$$

for arbitrary nonvanishing fiber vector with (ignorable coordinate) components ζ^r.

Having thus ensured the positivity of $\widetilde{\Sigma}$ as given by (7.9), we shall obtain an orthogonal transitivity theorem of the desired type whenever the differential in one or other version of the divergence formulas (7.10) vanishes, or at least has vanishing surface integral, over the limiting boundary surface of the region under consideration. The likelihood of being able to obtain such a result is enhanced by the possibility of exploiting the gauge dependence of ${}^{\flat}G_{rs}$ as defined in the terms of the $\Phi_r{}^X$ by (4.9), which can always be used to make it negative definite [in accordance with (7.11)] locally wherever g_{rs} is negative definite, but the possibility of choosing a gauge that does this globally in the region under consideration (i.e., from the higher-dimensional Kaluza–Klein point of view, the possibility of being able to choose a bundle section such that the surfaces of transitivity of the group action within it are everywhere timelike with respect to the induced metric) does not automatically follow, as is shown by the counterexample provided by the domains of outer communication of Riessner–Nordstrom black holes with charge to mass ratio sufficiently close to the critical value.[9]

ACKNOWLEDGMENTS

The author wishes to thank Gary Gibbons for raising the question that lead to this work, and to thank members of the U.C.S.B. Department of Physics, particularly James Hartle, Gary Horowitz, and Richard Woodard, for hospitality and informative conversations.

This research was supported by N. S. F. Grant No. PHY82-17853, supplemented by N.A.S.A., and by E.R. 176 of the C. N. R.S.

[1] W. Israel, Phys. Rev. **174**, 1776 (1967).
[2] W. Israel, Commun. Math. Phys. **8**, 245 (1968).
[3] A. Papapetrou, Ann. Inst. H. Poincaré A **4**, 83 (1966).
[4] B. Carter, J. Math. Phys. **10**, 70 (1969).
[5] S. W. Hawking, Commun. Math. Phys. **25**, 152 (1972); also in S. W. Hawking and G. F. R. Ellis, *The Large Scale Structure of Space-time* (Cambridge U. P., Cambridge, 1973).
[6] A. Lichnerowicz, *Theories Relativistes de la Gravitation et de l'Electromagnetism* (Masson, Paris, 1955).
[7] B. Carter, in *Black Holes (Les Houches 1972)*, edited by B. and C. DeWitt, (Gordon and Breach, New York, 1973), pp. 59–214.
[8] B. Carter, in *Gravitation in Astrophysics*, NATO ASI B **156** *(Cargèse 1986)*, edited by B. Carter and J. B. Hartle (Plenum, New York, 1987), pp. 63–122.
[9] B. Carter, "Electromagnetic ergoregions and conditions for staticity of non-rotating black holes," in *Proceedings of The 2nd Canadian Conference on General Relativity and Relativistic Astrophysics (Toronto, May, 1987)*, edited by A. Coley, C. Dyer, and B. Tupper (World Scientific, Singapore, 1987).
[10] P. Breitenlohner, D. Maison, and G. Gibbons, "4-dimensional black holes from Kaluza–Klein theory," D.A.M.T.P. preprint, Cambridge, 1986.
[11] T. Kaluza, Sitzung. Preuss. Akad. Wiss. Berlin, Math. Phys. K **1**, 966 (1921).
[12] O. Klein, Z. Phys. **37**, 895 (1926).
[13] B. DeWitt, in *Relativity, Groups, and Topology (Les Houches 1963)*, edited by B. DeWitt and C. DeWitt (Gordon and Breach, New York, 1964).
[14] R. Kerner, Ann. Inst. H. Poincaré **9**, 143 (1968).
[15] Y. M. Cho, J. Math. Phys. **16**, 2029 (1975).
[16] Y. M. Cho and P. G. O. Freund, Phys. Rev. D **12**, 1711 (1975).
[17] Y. M. Cho and P. S. Yang, Phys. Rev. D **12**, 3789 (1975).
[18] L. N. Chang, K. I. Macrae, and F. Mansoury, Phys. Rev. D **13**, 235 (1976).
[19] J. Scherk and J. Schwarz, Nucl. Phys. B **153**, 61 (1979).
[20] R. Geroch, J. Math. Phys. **12**, 918 (1971).
[21] F. J. Ernst, Phys. Rev. **167**, 1175 (1968).

The absence of small solutions of nonlinear field equations

Edward Malec
Institute of Physics, Jagellonian University, 30-059 Cracow, Reymonta 4, Poland

(Received 11 September 1986; accepted for publication 17 June 1987)

A simple criterion for the absence of small solutions of nonlinear field equations is proved. The key step in the proof consists of the use of energy Sobolev inequalities. Field-theoretic implications are shown.

I. INTRODUCTION

There is a tendency in the physical literature to look for those features of classical field theory (CFT) that correspond to certain properties that are expected to be inherent in quantum field theory (QFT). The two directions of intensive investigation are (i) the global Cauchy problem and classical field scattering[1] (the corresponding problem of the existence of S matrix in QFT is nothing more than a postulate in 3 + 1 Minkowski space); and (ii) the so-called classical color screening in non-Abelian gauge theory,[2] which is expected to mimic the confinement of quarks (still far from being proved in the framework of quantum chromodynamics).

The above list should be, in our opinion, extended to include the investigation of perturbative solutions of CFT. The incentive to do this comes from the fact that the perturbative approach is commonly used in QFT, where its validity cannot be checked, for obvious reasons.

This paper originated from an intention to fill the gap in our knowledge about the validity of the perturbation technique in CFT, but it comprises more general results. This work presents a simple tool for studying the existence problem of solutions small in a certain Banach norm to nonlinear field equations. The main results are contained in Theorems 1 and 2 (Sec. II). They state that under certain conditions (concerning integrability and nonlinearity) small solutions of nonlinear field equations in three space dimensions are absent. As a consequence we get also the absence of perturbative solutions.

Let us point out the distinguished role of $D = 3$ space dimensions; the above results are true only for $D = 3$ and $D > 3$. This is due to the fact that our approach relies on the use of kinetic energy Sobolev inequalities (see Sec. II below) which are known only in $D \geqslant 3$ space dimensions.

The obtained results are used to solve two outstanding problems in classical non-Abelian gauge theories. This is reported in Sec. III. We also apply our formalism to the nonlinear Klein–Gordon equation, to get the absence of perturbative solutions in cases when full solutions are know to exist.

II. MAIN RESULTS

The main result is the following.

Theorem 1: Let the equations of motion be

$$Af + N(|f|)f = 0, \qquad (1)$$

where f is a multicomponent complex valued field; A is a positive operator in the sense that

$$\int f^+ A f \, d^3x \geqslant c \int |\nabla f|^2 \, d^3x \geqslant 0, \quad c > 0; \qquad (2)$$

and $N(f)$ is homogeneous in the f's: $\|N(sf)\| = |s|^{p-2}\|N(f)\|$.[3] Assume that $f \in L_q(\mathbb{R}^3) \cap L_6(\mathbb{R}^3)$, $f \in L_2(\mathbb{R}^3)$; then $q = \frac{3}{2}(p-2)$, $p > 2$. Then the nonzero solutions of Eqs. (1) are absent provided that the L_q norm of f is sufficiently small.

Proof: Multiply Eq. (1) by f^+ (the Hermitian conjugate of f), integrate over all space $\mathbb{R}^3$, and rewrite it in the form

$$\int f^+ A f \, d^3x = -\int f^+ N(|f|) f \, d^3x. \qquad (3)$$

By the use of (2) and by the definition of the operator norm we get the following inequality:

$$c \int |\nabla f|^2 \, d^3x \leqslant \int \|N(|f|)\| |f|^2 \, d^3x. \qquad (4)$$

Now we will estimate the rhs of (4) by the use of Hölder inequalities:

$$\begin{aligned}\int \|N(f)\| \, |f|^2 \, d^3x &\leqslant \left(\int \|N(f)\|^{3/2} \, d^3x\right)^{2/3} \left(\int |f|^6 \, d^3x\right)^{2/6} \\ &\leqslant \left(\int \|N(f)\|^{3/2} \, d^3x\right)^{2/3} \frac{4}{3} \left(\int |\nabla f|^2 \, d^3x\right)^2. \qquad (5)\end{aligned}$$

The second inequaltiy follows from the Sobolev kinetic energy estimation, which, in three dimensions, takes the form[4]

$$\|f\|_{L_6} \leqslant [2/(3)^{1/2}] \|\nabla f\|_{L_2}. \qquad (6)$$

Now we employ the homogeneity of $N(f)$ as well as Minkowski and Hölder inequalities to get

$$\begin{aligned}\left(\int \|N(|f|)\|^{3/2} \, d^3x\right)^{2/3} &\leqslant c' \left(\int |f|^{3(p-2)/2} \, d^3x\right)^{2/3} \\ &= c' \|f\|_{L_q}^{(p-2)}, \qquad (7)\end{aligned}$$

where c' is a constant that depends only on coefficients appearing in $N(f)$.

Using (7) and (5) to estimate the rhs of (4) we eventually arrive at

$$c \int |\nabla f|^2 \, d^3x \leqslant c' \frac{4}{3} \|f\|_{L_q}^{p-2} \int |\nabla f|^2 \, d^3x; \qquad (8)$$

thus for

$$\|f\|_{L_q}^{p-2} < 3c/4c',$$

the only possibility is $\nabla f = 0$, that is, $f = 0$ [since $f \in L_q(\mathbb{R}^3)$]. This concludes our proof.

0022-2488/88/010235-03$02.50

For nonlinear systems a rescaling of fields allows us to introduce a parameter into the equations of motion. For instance, the rescaling $f \to fs^{-1/(p-2)}$ in Eq. (1) gives equations with a "coupling constant" s:

$$Af + sN(|f|)f = 0. \tag{9}$$

Therefore we get the following as a direct conclusion from Theorem 1.

Theorem 2: Let the assumptions concerning f and A be as in Theorem 1. Then nonzero solutions of Eqs. (9) are absent provided that the constant t, defined by

$$t = \lim_{s \to 0} s\|f\|_{L_q}^{p-2}, \tag{10}$$

is small enough.

Corollary: Under the conditions of Theorem 2, Eqs. (9) have no nontrivial solutions analytic at $s = 0$.

Proof of the Corollary: Notice that for perturbative solutions $t = 0$ and use Theorem 2.

Remark 1: The above results are immediately extended to equations with polynomial interaction terms $N = \Sigma_i N_i$, where the N_i are characterized by different degrees of homogeneity q_i,

$$\|N_i(sf)\| = |s|^{q_i}\|N(f)\|, \quad q_i > 0.$$

Remark 2: Similar results hold in D-space dimensions for $D > 3$, but, of course, under different integrability conditions. For needed Sobolev estimations see, e.g., Ref. 4. It is interesting to notice that for $D < 2$ the above approach does not work, since there is no kinetic energy estimation (6).

Remark 3: A more sophisticated investigation of a nonlinear elliptic equation would give results stronger than those above on the absence of nontrivial solutions. Theorem 2.1 in Ref. 5, for instance, states the *global absence* of solutions to certain classes of the nonlinear Klein–Gordon equation. Stronger results, however, require a more complicated machinery (e.g., conservation laws and other useful identities[5]), which happens to be strongly dependent on the specific equation. Therefore the range of their validity is narrower in comparison with our Theorem 1. To exemplify the last statement, let us point out that our Theorem 1 guarantees the absence of small solutions in the case D of Example 2 in Ref. 5, in which the global methods say nothing.

III. APPLICATIONS

Below we briefly report on the field-theoretic applications of the results obtained earlier. The first two examples are described in more detail in Ref. 6.

(a) *Screening solutions*[2] seem to be absent in self-contained Yang–Mills–matter field theory. Because of "no go" theorems (the simplest one is contained in Ref. 7) the only reasonable candidates for revealing the classical color screening are theories with dynamical sources carried by fermion fields. The crucial point in the proof is the observation that screening solutions of Sikivie and Weiss[2] (which were found for fixed external sources) are characterized by the quantity $t = 0$ [see (10) for the definition of t] and that for the coupled Yang–Mills–Dirac equations an inequality of the type (8) can be obtained.

(b) *Perturbative solutions* of self-consistent Yang–Mills–Dirac equations are absent. The reasoning is similar to that in (a). Let us point out that several authors have been looking for such solutions for external fixed sources.[8]

(c) For the *nonlinear Klein-Gordon equation*

$$-\Delta f + m^2 f + s|f|^{p-2} f = 0, \tag{11}$$

the nonzero solutions satisfying the same conditions as in Theorem 2 and in addition $|f\Delta f| = o(r^{-2})$, $r \gg 1$ (this ensures that $-\int f^{+}\Delta f\, d^3x \geqslant 0$) are absent if $t < \frac{3}{4}$. Notice that for $p > 5$ and $s < 0$ solutions of Eq. (11) do exist.[9] They cannot be obtained perturbatively. This is a good occasion to show the superiority of our approach over the formal perturbation expansion of f, which in this simple case also gives the absence of perturbation solutions. The point is that our method gives a precise bound on $t = \lim_{s\to 0} s\|f\|_{L_q}^{p-2}$, from which we deduce two facts. First, the nonzero solution of Eq. (11) should be singular at $s = 0$ and, second, the coefficient at the leading term (in the expansion of f in a series of powers of s) should be sufficiently large.

Let us end with the remark that for complicated systems of equations our approach is much simpler than the standard one. There are also cases in which the formal perturbation expansion could lead to the wrong conclusion that perturbation solutions do exist (that is, the solution of linearized equations is tangent to a full solution of nonlinear equations). Here is an example:

$$\begin{pmatrix} \Delta & l \\ l & \Delta \end{pmatrix}\begin{pmatrix} u \\ v \end{pmatrix} + sN(u,v) = 0. \tag{12}$$

In (12), $n(u,v)$ is nonlinear and satisfies the conditions of Theorem 1 while l is an imaginary parameter.

Theorem 1 excludes the existence of small solutions of (12) (under suitable conditions on the falloff at the spatial infinity) while the standard analysis could suggest the existence of perturbation solutions (the equations linearized at $u = v = 0$ possess nonzero solutions for imaginary l), regardless of the form of nonlinearity in $N(u,v)$. It is easy to find N's such that Eqs. (12) have no full solutions, in accordance with Theorem 1 and contrary to the standard perturbation analysis.

ACKNOWLEDGMENT

The author was supported in part under the project 01.03.

[1] For example, W. A. Strauss, J. Funct. Anal. **41**, 110 (1981); D. Eardley, and V. Moncrief, Commun. Math. Phys. **83**, 171 (1982); I. Segal, J. Funct. Anal. **33**, 175 (1979); J. Ginibre and G. Velo, Commun. Math. Phys. **82**, 1 (1981); R. T. Glassey and W. A. Strauss, *ibid.* **89**, 465 (1983); see, also, references therein.

[2] P. Sikivie and N. Weiss, Phys. Rev. Lett. **40**, 1411 (1978); Phys. Rev. D **18**, 489 (1978); C. H. Lai and C. H. Oh, *ibid.* **29**, 1805 (1984), and references therein.

[3] $\|N\|$ denotes the following norm of the finite-dimensional matrix N: $\|N\| = \sup(|f^{+}Nf|/|f|^2)$.

[4] M. S. Berger, *Nonlinearity and Functional Analysis* (Academic, New York, 1977). The refined version of Sobolev inequality is contained in E. H. Lieb, Rev. Mod. Phys. **48**, 553 (1976) and G. Rosen, SIAM J. Appl. Math. **21**, 30 (1971).

[5] W. A. Strauss, in *Invariant Wave Equations, Lecture Notes in Physics*, Vol. 73, edited by G. Velo and A. Wightman (Springer, Berlin, 1978), Secs. 2 and 3.

[6] A. Górski, and E. Malec, "On coupling constant dependence of gauge fields," TPJU preprint 12, 1986.

[7] S. Deser, Phys. Lett. B **64**, 463 (1976).

[8] For example, R. Jackiw, L. Jacobs, and C. Rebbi, Phys. Rev. D **20**, 474 (1979).

[9] J. Shatah and W. Strauss, Commun. Math. Phys. **100**, 173 (1985), and references therein.

Nonlocal symmetries and Bäcklund transformations for the self-dual Yang–Mills system

C. J. Papachristou and B. Kent Harrison
Department of Physics and Astronomy, Brigham Young University, Provo, Utah 84602

(Received 15 May 1987; accepted for publication 2 September 1987)

The observation is made that generalized evolutionary isovectors of the self-dual Yang–Mills equation, obtained by "verticalization" of the geometrical isovectors derived in a previous paper [J. Math. Phys. **28**, 1261 (1987)], generate Bäcklund transformations for the self-dual system. In particular, new Bäcklund transformations are obtained by "verticalizing" the generators of point transformations on the solution manifold. A geometric ansatz for the derivation of such (generally nonlocal) symmetries is proposed.

I. INTRODUCTION

In previous papers[1,2] the authors have discussed isovector[3] techniques for partial differential equations (PDE's) associated with vector-valued differential forms. It was mentioned[2] that such a PDE (or system of PDE's) defines, through its solutions, sections of a vector bundle over the solution manifold. This manifold serves as a base space, while the fibers are isomorphic to some vector space (or Lie algebra).

In Ref. 1 the isovector approach was employed to derive point symmetries for the self-dual Yang–Mills (SDYM) equation in its so-called J formulation[4] (this is mathematically different from the usual formulation in which the self-duality condition is directly written in covariant form). The system was represented by three gl(N,C)-valued four-forms in seven variables. Since these forms generated a differential ideal by themselves, we did not include the integrability condition of the system in the ideal (indeed, such an inclusion can be seen to be superfluous for the purpose of deriving *point* transformations).

Calculation of the isovectors gave a nine parameter group of transformations on the base space, together with a set of infinitesimal internal transformations in the fiber space in which the SDYM fields have values. It was then observed that the internal symmetries were related to parametric Bäcklund transformations (BT's) for the SDYM equation. In particular, a well-known[5,6] BT was recovered.

With the observation that internal symmetries are generated by *evolutionary*[7] (i.e., "vertical") vector fields (EVF's), it is natural to inquire for other EVF's that may generate BT's for the SDYM system. Such fields cannot be sought, of course, among the "geometrical" symmetries found in Ref. 1. The most accessible nongeometrical (i.e., nonlocal) symmetries at our disposal are those generated by the *evolutionary representatives*[7] of the nine generators of coordinate transformations mentioned previously. In Sec. III we establish the generalized isovector property of these EVF's by examining the effect of the corresponding Lie derivatives on the original ideal of the three four-forms. It is found that these Lie derivatives map this ideal into a larger ideal comprising the original system, its integrability conditions, and certain prolongations[8] of all of the above. The emergence of a prolonged ideal was to be expected since we are now dealing with Lie–Bäcklund-type symmetries. On the other hand, the appearance of the integrability conditions as an inseparable part of the system is quite interesting, considering the passive role these conditions played in the derivation of point symmetries.

In Sec. IV we construct the infinitesimal parametric nonlocal transformations generated by the aforementioned nine EVF's. It is seen that, by letting the transformation parameters be considered finite, rather than infinitesimal, the transformations of the prolongation[8] variables become BT's for the SDYM equation. This observation constitutes a further indication of the intimate connection between symmetry and integrability aspects[9] of nonlinear systems, and, in particular, of the SDYM system.[1,10]

To make the paper as self-contained as possible, we review in the next section some of the results of Ref. 1 that will be needed for the present treatment.

II. GEOMETRICAL SYMMETRIES OF THE SDYM SYSTEM

As in Ref. 1, we write the SDYM equation (a second-order nonlinear PDE) as a set of first-order PDE's:

$$B^1_{\bar y} + B^2_{\bar z} = 0, \qquad (2.1a)$$

$$B^1 = J^{-1}J_y, \qquad (2.1b)$$

$$B^2 = J^{-1}J_z, \qquad (2.1c)$$

where the subscripts denote partial differentiation (partial derivatives will occasionally be used). The $y, z, \bar y, \bar z$, collectively denoted by x^μ ($\mu = 1,2,3,4$), are four complex coordinates,[1,4,5] while J is assumed to have values in gl(N,C). Loosely speaking, the B^1 and B^2 are "prolongation variables" for the system. The integrability condition $J_{yz} = J_{zy}$ yields

$$B^2_y - B^1_z + [B^1, B^2] = 0. \qquad (2.2)$$

The system (2.1) can be represented by a set of three gl(N,C)-valued four-forms:

$$\begin{aligned} \gamma_1 &= dy\, dz\, dB^1\, d\bar z + dy\, dz\, d\bar y\, dB^2, \\ \gamma_2 &= dJ\, dz\, d\bar y\, d\bar z - JB^1\, dy\, dz\, d\bar y\, d\bar z, \qquad (2.3) \\ \gamma_3 &= dy\, dJ\, d\bar y\, d\bar z - JB^2\, dy\, dz\, d\bar y\, d\bar z. \end{aligned}$$

0022-2488/88/010238-06$02.50

These forms generate a differential ideal. Indeed,

$$d\gamma_1 = 0, \quad d\gamma_2 = \gamma_2\, dy\, B^1 + J\, d\bar{z}\, \gamma_1,$$
$$d\gamma_3 = \gamma_3\, dz\, B^2 + J\, d\bar{z}\, \gamma_1.$$

The geometrical (local) symmetries of the system are generated by vector fields of the form[1]

$$V = \xi^{\mu}(x^{\nu})\frac{\partial}{\partial x^{\mu}} + G(x^{\nu},J)\frac{\partial}{\partial J} + A^{i}(x^{\nu},B^{k})\frac{\partial}{\partial B^{i}}, \qquad (2.4)$$

where the usual summation convention is assumed over repeated indices. The ξ^{μ} ($\mu = 1,...,4$) are scalars, whereas the G and A^i ($i = 1,2$) are gl(N,C) valued. The $\partial/\partial J$ and $\partial/\partial B^i$ are *formal* operators only—i.e., no differentiations are actually performed. The symmetry property is expressed by the requirement that the Lie derivative with respect to V leave the ideal of the γ_k invariant. Formally,

$$\underset{V}{\pounds}\, \gamma_i = b_{i}^{k}\gamma_k + \Lambda_i^k\gamma_k + \gamma_k M_{i}^{k} \qquad (2.5)$$

($i = 1,2,3$), where the b_{i}^{k} are scalars, whereas the Λ_i^k and M_{i}^{k} are gl(N,C)-valued zero-forms. The calculation of V from Eq. (2.5) becomes possible if we make the ansatz that the coefficients of expansion in this equation depend only on the x^{μ}. As we will see shortly, this condition is violated in the case of nongeometrical symmetries.

As was seen in Ref. 1, the vector V is parametrized by nine (complex) parameters, and depends on three arbitrary functions. The nine parameters correspond to transformations on the base space. These transformations are generated by the following independent vector fields, which are written here in unprolonged form (i.e., without the terms in $\partial/\partial B^i$):

$$V_{\mu} = -\frac{\partial}{\partial x^{\mu}}, \quad \mu = 1,2,3,4, \quad V_5 = -y\frac{\partial}{\partial y} - \bar{z}\frac{\partial}{\partial \bar{z}},$$
$$V_6 = -z\frac{\partial}{\partial z} - \bar{y}\frac{\partial}{\partial \bar{y}}, \quad V_7 = -z\frac{\partial}{\partial y} + \bar{y}\frac{\partial}{\partial \bar{z}}, \qquad (2.6)$$
$$V_8 = -y\frac{\partial}{\partial z} + \bar{z}\frac{\partial}{\partial \bar{y}}, \quad V_9 = -\bar{y}\frac{\partial}{\partial \bar{y}} - \bar{z}\frac{\partial}{\partial \bar{z}}.$$

Internal symmetries are generated by EVF's with components

$$G = \epsilon(\bar{y},\bar{z})J + \Lambda(\bar{y},\bar{z})J + JM(y,z),$$
$$A^1 = -[M(y,z),B^1] + M_y,$$
$$A^2 = -[M(y,z),B^2] + M_z,$$

where ϵ is a scalar function, while Λ and M are gl(N,C)-valued functions. The symmetries in ϵ and M combine to give BT's for SDYM.[1] The symmetry in Λ yields a BT that is less interesting, since it merely consists of $\delta(J^{-1}J_y) = 0$, $\delta(J^{-1}J_z) = 0$ (i.e., $J'^{-1}J'_y = J^{-1}J_y$, etc.).

III. EVOLUTIONARY ISOVECTORS FOR FIRST-ORDER GENERALIZED SYMMETRIES

A. Nongeometrical vectors and prolongation forms

We now relax the geometrical requirement and seek generalized[7] symmetries of SDYM. We confine our attention to first-order symmetries generated by *evolutionary* vector fields (i.e., vector fields with vanishing projection on the base space) of the form

$$V = Q\frac{\partial}{\partial J} + R^1\frac{\partial}{\partial B^1} + R^2\frac{\partial}{\partial B^2}, \qquad (3.1)$$

where Q may depend on x^{μ}, J, and $J_{\mu} \equiv \partial_{\mu}J$. With the definitions

$$J_y = JB^1, \quad J_z = JB^2, \qquad (3.2a)$$
$$J_{\bar{y}} = E^1, \quad J_{\bar{z}} = E^2, \qquad (3.2b)$$

we have that

$$Q = Q(x^{\mu},J,B^1,B^2,E^1,E^2). \qquad (3.3)$$

[The reader may be concerned with the appearance of several noncommuting variables in the functional dependence of Q. However, as Q is calculated (see below), no ambiguity in the order of these variables arises.]

It is easily seen [see Eqs. (3.11) and (3.12)] that the R^1 and R^2 depend, collectively, on the additional variables $B_{\mu}^{1} \equiv \partial_{\mu}B^1$ and $B_{\mu}^{2} \equiv \partial_{\mu}B^2$. The variables $B_{\bar{z}}^{2}$ and B_{z}^{1} can be eliminated from the problem by using the field equation (2.1a) and the integrability condition (2.2), respectively:

$$B_{\bar{z}}^{2} = -B_{\bar{y}}^{1}, \qquad (3.4)$$
$$B_{z}^{1} = B_{y}^{2} + [B^1,B^2]. \qquad (3.5)$$

Thus we are left with the variables

$$B_{y}^{1} = C^1, \quad B_{\bar{y}}^{1} = C^2, \quad B_{\bar{z}}^{1} = C^3,$$
$$B_{y}^{2} = C^4, \quad B_{z}^{2} = C^5, \quad B_{\bar{y}}^{2} = C^6. \qquad (3.6)$$

Equations (3.2) and (3.6) each admit six integrability conditions, thus a total of 12 such conditions can be written [including the one given by Eq. (2.2)].

Let us consider the basic system (2.1), together with its integrability condition (2.2). We prolong Eqs. (2.1a) and (2.2) in the usual way by taking the derivatives with respect to the x^{μ}. By convention, only those prolongations defined *within the variables at our disposal* are considered. Specifically, we can construct the y, z, and $\bar{y}$ prolongations of Eq. (2.1a), and the y and $\bar{y}$ prolongations of Eq. (2.2):

$$C_{\bar{y}}^{1} + C_{\bar{z}}^{4} = 0, \qquad (3.7a)$$
$$C_{z}^{2} + C_{\bar{z}}^{5} = 0, \qquad (3.7b)$$
$$C_{\bar{y}}^{2} + C_{\bar{z}}^{6} = 0; \qquad (3.7c)$$
$$C_{y}^{4} - C_{z}^{1} + [C^1,B^2] + [B^1,C^4] = 0, \qquad (3.8a)$$
$$C_{y}^{6} - C_{z}^{2} + [C^2,B^2] + [B^1,C^6] = 0. \qquad (3.8b)$$

[Note that, in a sense, the prolongations of Eqs. (2.1b) and (2.1c) are contained in Eq. (3.6).]

We now express our equations in terms of differential forms. Thus we define 28 four-forms corresponding, successively, to Eqs. (2.1), (2.2), (3.7), (3.8), (3.6) and its integrability conditions, and Eq. (3.2b), and the five remaining integrability conditions of (3.2) (we put $\widetilde{\omega} \equiv dy\, dz\, d\bar{y}\, d\bar{z}$):

$$\gamma_1 = dy\, dz\, dB^1\, d\bar{z} + dy\, dz\, d\bar{y}\, dB^2,$$
$$\gamma_2 = dJ\, dz\, d\bar{y}\, d\bar{z} - JB^1\widetilde{\omega},$$
$$\gamma_3 = dy\, dJ\, d\bar{y}\, d\bar{z} - JB^2\widetilde{\omega},$$
$$\gamma_4 = dB^2\, dz\, d\bar{y}\, d\bar{z} - dy\, dB^1\, d\bar{y}\, d\bar{z} + [B^1,B^2]\widetilde{\omega},$$
$$\gamma_5 = dy\, dz\, dC^1\, d\bar{z} + dy\, dz\, d\bar{y}\, dC^4,$$
$$\gamma_6 = dy\, dC^2\, d\bar{y}\, d\bar{z} + dy\, dz\, d\bar{y}\, dC^5,$$

$$
\begin{aligned}
\gamma_7 &= dy\, dz\, dC^2\, d\bar{z} + dy\, dz\, d\bar{y}\, dC^6,\\
\gamma_8 &= dC^4\, dz\, d\bar{y}\, d\bar{z} - dy\, dC^1\, d\bar{y}\, d\bar{z}\\
&\quad + ([C^1,B^2] + [B^1,C^4])\tilde{\omega},\\
\gamma_9 &= dC^6\, dz\, d\bar{y}\, d\bar{z} - dy\, dC^2\, d\bar{y}\, d\bar{z}\\
&\quad + ([C^2,B^2] + [B^1,C^6])\tilde{\omega},\\
\gamma_{10} &= dB^1\, dz\, d\bar{y}\, d\bar{z} - C^1\tilde{\omega},\\
\gamma_{11} &= dy\, dz\, dB^1\, d\bar{z} - C^2\tilde{\omega},\\
\gamma_{12} &= dy\, dz\, d\bar{y}\, dB^1 - C^3\tilde{\omega},\\
\gamma_{13} &= dB^2\, dz\, d\bar{y}\, d\bar{z} - C^4\tilde{\omega},\\
\gamma_{14} &= dy\, dB^2\, d\bar{y}\, d\bar{z} - C^5\tilde{\omega},\\
\gamma_{15} &= dy\, dz\, dB^2\, d\bar{z} - C^6\tilde{\omega},\\
\gamma_{16} &= dC^2\, dz\, d\bar{y}\, d\bar{z} - dy\, dz\, dC^1\, d\bar{z},\\
\gamma_{17} &= dC^3\, dz\, d\bar{y}\, d\bar{z} - dy\, dz\, d\bar{y}\, dC^1,\\
\gamma_{18} &= dy\, dz\, dC^3\, d\bar{z} - dy\, dz\, d\bar{y}\, dC^2,\\
\gamma_{19} &= dC^5\, dz\, d\bar{y}\, d\bar{z} - dy\, dC^4\, d\bar{y}\, d\bar{z},\\
\gamma_{20} &= dC^6\, dz\, d\bar{y}\, d\bar{z} - dy\, dz\, dC^4\, d\bar{z},\\
\gamma_{21} &= dy\, dC^6\, d\bar{y}\, d\bar{z} - dy\, dz\, dC^5\, d\bar{z},\\
\gamma_{22} &= dy\, dz\, dJ\, d\bar{z} - E^1\tilde{\omega},\\
\gamma_{23} &= dy\, dz\, d\bar{y}\, dJ - E^2\tilde{\omega},\\
\gamma_{24} &= dy\, dz\, dE^2\, d\bar{z} - dy\, dz\, d\bar{y}\, dE^1,\\
\gamma_{25} &= dE^1\, dz\, d\bar{y}\, d\bar{z} - dy\, dz\, d(JB^1)\, d\bar{z},\\
\gamma_{26} &= dE^2\, dz\, d\bar{y}\, d\bar{z} - dy\, dz\, d\bar{y}\, d(JB^1),\\
\gamma_{27} &= dy\, dE^1\, d\bar{y}\, d\bar{z} - dy\, dz\, d(JB^2)\, d\bar{z},\\
\gamma_{28} &= dy\, dE^2\, d\bar{y}\, d\bar{z} - dy\, dz\, d\bar{y}\, d(JB^2)\,.
\end{aligned}
\tag{3.9}
$$

B. Generalized isovectors

The postulate (2.5), used to derive symmetries of the system, can now be generalized[11] by requiring that the Lie derivative with respect to a generalized EVF of the form (3.1) map the original ideal $\{\gamma_1,\gamma_2,\gamma_3\}$ into the prolonged ideal $\{\gamma_1,...,\gamma_{28}\}$. [We can see that such a generalization is necessary because the Lie derivative of the γ_i $(i = 1,2,3)$ will yield variables which appear only in the prolonged forms γ_α.] Formally,

$$\underset{V}{\pounds}\, \gamma_i = b_i^\alpha \gamma_\alpha + \Lambda_i^\alpha \gamma_\alpha + \gamma_\alpha M_i^\alpha, \tag{3.10}$$

where $i = 1,2,3$, as before, but now the index α runs from 1 to 28. The coefficients Λ_i^α and M_i^α are no longer required to be independent of the internal variables J, B^1, and B^2.

One would like, of course, to solve Eq. (3.10) for the components Q, R^1, and R^2 of V and thus obtain independent first-order generalized symmetries of SDYM. The computations, however, are now of great complexity due to the presence of a very large set of variables and forms. Rather than solving Eq. (3.10) directly, we will instead construct certain solutions by utilizing the coordinate (point) symmetries of SDYM, as these are expressed by the vector fields of Eq. (2.6).

First of all, it is seen from Eqs. (3.9) and (3.10) that some of the forms of the prolonged ideal will not occur in the expansion of the Lie derivative in Eq. (3.10). Indeed, by comparing similar terms and taking into account Eq. (3.3), it can be shown that the forms γ_8, γ_{16}, γ_{17}, γ_{19}, and γ_{21} make no contribution and thus can be eliminated. This leaves us with a somewhat smaller ideal of 23 forms.

Second, since the B^1 and B^2 are prolongation variables, the R^1 and R^2 of Eq. (3.1) are expressible in terms of Q. Again, this can be done by comparing similar terms in Eq. (3.10), using the "internal exterior derivative" (introduced in Refs. 1 and 2) whenever necessary.

There is, however, an easier way to do this: Let us recall that the vector field V, in the form (3.1), defines an infinitesimal "motion" in the space of J, B^1, and B^2. Thus, if δt is an infinitesimal parameter, $\delta J = Q\delta t$. Furthermore, if we regard t as a variable parametrizing an integral curve of V, then $Q = \delta J/\delta t$ and

$$
\begin{aligned}
R^1 = \frac{\delta B^1}{\delta t} &= -J^{-1}\frac{\delta J}{\delta t}J^{-1}D_y J + J^{-1}\frac{\delta}{\delta t}(D_y J)\\
&= -J^{-1}QB^1 + J^{-1}D_y Q
\end{aligned}
\tag{3.11}
$$

and similarly

$$R^2 = \frac{\delta B^2}{\delta t} = -J^{-1}QB^2 + J^{-1}D_z Q, \tag{3.12}$$

where the D_y and D_z are total derivatives.[12] [In general, total derivatives must be defined consistently with Eqs. (3.2) and (3.4)–(3.6). Thus, on functions of the form (3.3),

$$
\begin{aligned}
D_y &= \frac{\partial}{\partial y} + JB^1\frac{\partial}{\partial J} + C^1\frac{\partial}{\partial B^1} + C^4\frac{\partial}{\partial B^2}\\
&\quad + (E^1B^1 + JC^2)\frac{\partial}{\partial E^1} + (E^2B^1 + JC^3)\frac{\partial}{\partial E^2},\\
D_z &= \frac{\partial}{\partial z} + JB^2\frac{\partial}{\partial J} + (C^4 + [B^1,B^2])\frac{\partial}{\partial B^1} + C^5\frac{\partial}{\partial B^2}\\
&\quad + (E^1B^2 + JC^6)\frac{\partial}{\partial E^1} + (E^2B^2 - JC^2)\frac{\partial}{\partial E^2}.
\end{aligned}
$$

Note that D_μ, like ∂_μ, is a *derivation*, i.e., it satisfies the Leibniz rule.]

Finally, we need to define the *evolutionary representative*[7] (or verticalization) of a "horizontal" vector field. The following definition is pertinent to the SDYM case but can be easily generalized: Given a vector field of the (unprolonged) form

$$V_h = \xi^\mu(x^\nu)\frac{\partial}{\partial x^\mu}, \tag{3.13}$$

one can construct an EVF (in prolonged form)

$$V_q = Q\frac{\partial}{\partial J} + R^1\frac{\partial}{\partial B^1} + R^2\frac{\partial}{\partial B^2} \tag{3.14}$$

such that

$$Q = -\xi^\mu D_\mu J, \tag{3.15}$$

where $D_y J = JB^1$, $D_z J = JB^2$, $D_{\bar{y}} J = E^1$, $D_{\bar{z}} J = E^2$, and where the components R^1 and R^2 are related to Q as in Eqs. (3.11) and (3.12). The nongeometrical vector field V_q is called the evolutionary representative of the geometrical field V_h.[13] The reason for this particular definition is that, if V_h is a symmetry, it can be shown[7] that V_q is also a symmetry.

We are now in a position to construct the (prolonged) evolutionary representatives of the nine vector fields given in Eq. (2.6). The symbol $\overline{V}_k$ will denote the representative of V_k:

$$
\begin{aligned}
\overline{V}_1 &= JB^1\frac{\partial}{\partial J}+C^1\frac{\partial}{\partial B^1}+C^4\frac{\partial}{\partial B^2},\\
\overline{V}_2 &= JB^2\frac{\partial}{\partial J}+(C^4+[B^1,B^2])\frac{\partial}{\partial B^1}+C^5\frac{\partial}{\partial B^2},\\
\overline{V}_3 &= E^1\frac{\partial}{\partial J}+C^2\frac{\partial}{\partial B^1}+C^6\frac{\partial}{\partial B^2},\\
\overline{V}_4 &= E^2\frac{\partial}{\partial J}+C^3\frac{\partial}{\partial B^1}-C^2\frac{\partial}{\partial B^2},\\
\overline{V}_5 &= (yJB^1+\bar{z}E^2)\frac{\partial}{\partial J}+(B^1+yC^1+\bar{z}C^3)\frac{\partial}{\partial B^1}\\
&\quad+(yC^4-\bar{z}C^2)\frac{\partial}{\partial B^2},\\
\overline{V}_6 &= (zJB^2+\bar{y}E^1)\frac{\partial}{\partial J}+(zC^4+z[B^1,B^2]+\bar{y}C^2)\\
&\quad\times\frac{\partial}{\partial B^1}+(B^2+zC^5+\bar{y}C^6)\frac{\partial}{\partial B^2},\qquad (3.16)\\
\overline{V}_7 &= (zJB^1-\bar{y}E^2)\frac{\partial}{\partial J}+(zC^1-\bar{y}C^3)\frac{\partial}{\partial B^1}\\
&\quad+(B^1+zC^4+\bar{y}C^2)\frac{\partial}{\partial B^2},\\
\overline{V}_8 &= (yJB^2-\bar{z}E^1)\frac{\partial}{\partial J}\\
&\quad+(B^2+yC^4+y[B^1,B^2]-\bar{z}C^2)\frac{\partial}{\partial B^1}\\
&\quad+(yC^5-\bar{z}C^6)\frac{\partial}{\partial B^2},\\
\overline{V}_9 &= (\bar{y}E^1+\bar{z}E^2)\frac{\partial}{\partial J}+(\bar{y}C^2+\bar{z}C^3)\frac{\partial}{\partial B^1}\\
&\quad+(\bar{y}C^6-\bar{z}C^2)\frac{\partial}{\partial B^2}.
\end{aligned}
$$

The consistency of these expressions with the geometrical derivation of symmetries, as this is expressed in the prescription (3.10), may now be seen. Direct substitution into Eq. (3.10) shows that the above EVF's are generalized isovectors for the SDYM system (details are deferred to Appendix A). Our search for other generalized symmetries, of the same order or higher, has not been successful so far. In particular, one can see that the (generalized) Lie brackets[7] of the known symmetries do not produce new symmetries, contrary to what one might have hoped (this point is more easily verified by using true jet-space variables x^μ, J, J_μ, and $J_{\mu\nu}$).

IV. INFINITESIMAL NONLOCAL SYMMETRIES AND BÄCKLUND TRANSFORMATIONS

By using the EVF's (3.16) and the definitions (3.2), the following infinitesimal parametric nonlocal transformations of the J function are constructed[7] (λ is an infinitesimal parameter):

$$
\begin{aligned}
&\delta_\mu J=\lambda\,\partial_\mu J\equiv\lambda J_\mu,\ \mu=1,2,3,4,\ \delta_5 J=\lambda(yJ_y+\bar{z}J_{\bar{z}}),\\
&\delta_6 J=\lambda(zJ_z+\bar{y}J_{\bar{y}}),\quad \delta_7 J=\lambda(zJ_y-\bar{y}J_{\bar{z}}),\qquad (4.1)\\
&\delta_8 J=\lambda(yJ_z-\bar{z}J_{\bar{y}}),\quad \delta_9 J=\lambda(\bar{y}J_{\bar{y}}+\bar{z}J_{\bar{z}}).
\end{aligned}
$$

We can also construct the corresponding transformations for the variables $B^1=J^{-1}J_y$ and $B^2=J^{-1}J_z$. We will initially regard these transformations as general nonlocal symmetries generated by the EVF's (3.16) and independent of the SDYM equation. This means that one is allowed to make replacements in the components of the EVF's according to the definitions (3.2) and (3.6) and the integrability condition (3.5), but one may not use the equation of motion (3.4) (thus the apparent asymmetry of the equations below). The transformation equations are

$$
\begin{aligned}
\delta_1(J^{-1}J_y)&=\lambda\,\partial_y(J^{-1}J_y),\\
\delta_1(J^{-1}J_z)&=\lambda\,\partial_y(J^{-1}J_z),\\
\delta_2(J^{-1}J_y)&=\lambda\,\partial_z(J^{-1}J_y),\\
\delta_2(J^{-1}J_z)&=\lambda\,\partial_z(J^{-1}J_z),\\
\delta_3(J^{-1}J_y)&=\lambda\,\partial_{\bar{y}}(J^{-1}J_y),\\
\delta_3(J^{-1}J_z)&=\lambda\,\partial_{\bar{y}}(J^{-1}J_z),\\
\delta_4(J^{-1}J_y)&=\lambda\,\partial_{\bar{z}}(J^{-1}J_y),\\
\delta_4(J^{-1}J_z)&=-\lambda\,\partial_{\bar{y}}(J^{-1}J_y),\\
\delta_5(J^{-1}J_y)&=\lambda\,[J^{-1}J_y+y\,\partial_y(J^{-1}J_y)+\bar{z}\,\partial_{\bar{z}}(J^{-1}J_y)],\\
\delta_5(J^{-1}J_z)&=\lambda\,[y\,\partial_y(J^{-1}J_z)-\bar{z}\,\partial_{\bar{y}}(J^{-1}J_y)],\qquad (4.2)\\
\delta_6(J^{-1}J_y)&=\lambda\,[z\,\partial_z(J^{-1}J_y)+\bar{y}\,\partial_{\bar{y}}(J^{-1}J_y)],\\
\delta_6(J^{-1}J_z)&=\lambda\,[J^{-1}J_z+z\,\partial_z(J^{-1}J_z)+\bar{y}\,\partial_{\bar{y}}(J^{-1}J_z)],\\
\delta_7(J^{-1}J_y)&=\lambda\,[z\,\partial_y(J^{-1}J_y)-\bar{y}\,\partial_{\bar{z}}(J^{-1}J_y)],\\
\delta_7(J^{-1}J_z)&=\lambda\,[J^{-1}J_y+z\,\partial_y(J^{-1}J_z)+\bar{y}\,\partial_{\bar{y}}(J^{-1}J_y)],\\
\delta_8(J^{-1}J_y)&=\lambda\,[J^{-1}J_z+y\,\partial_z(J^{-1}J_y)-\bar{z}\,\partial_{\bar{y}}(J^{-1}J_y)],\\
\delta_8(J^{-1}J_z)&=\lambda\,[y\,\partial_{\dot{z}}(J^{-1}J_z)-\bar{z}\,\partial_{\bar{y}}(J^{-1}J_z)],\\
\delta_9(J^{-1}J_y)&=\lambda\,[\bar{y}\,\partial_{\bar{y}}(J^{-1}J_y)+\bar{z}\,\partial_{\bar{z}}(J^{-1}J_y)],\\
\delta_9(J^{-1}J_z)&=\lambda\,[\bar{y}\,\partial_{\bar{y}}(J^{-1}J_z)-\bar{z}\,\partial_{\bar{y}}(J^{-1}J_y)].
\end{aligned}
$$

We now observe that the above nonlocal transformations of the prolongation variables bear an interesting property: Suppose that we let the parameters λ become *finite* in each case (λ stands for nine different independent parameters). In this case the left-hand sides of Eq. (4.2) become *finite* differences:

$$
\begin{aligned}
\delta_k(J^{-1}J_y)&\equiv J'^{-1}J'_y-J^{-1}J_y,\\
\delta_k(J^{-1}J_z)&\equiv J'^{-1}J'_z-J^{-1}J_z.
\end{aligned}\qquad (4.3)
$$

We thus obtain nine pairs of independent parametric equations. Cross-differentiation of each pair with respect to $\bar{y}$ and $\bar{z}$, and use of the various integrability conditions, will then reveal that all nine pairs are parametric Bäcklund transformations for the SDYM equation, which J is now required to satisfy. Specifically, if J is a solution of

$$\partial_{\bar y}(J^{-1}J_y) + \partial_{\bar z}(J^{-1}J_z) = 0, \tag{4.4}$$

then so is J', which is related to J by Eqs. (4.2) and (4.3). A more detailed proof of this is given in Appendix B.

We remark that one could obtain a more symmetric set of equations than Eq. (4.2) by allowing the components of the EVF's (3.16) to be evaluated on solutions of the SDYM equation, i.e., by using Eq. (3.4) to reintroduce $B^2_{\bar z}$ into the problem. The reader is invited to construct this alternate set of BT's.

In order for the BT's described in Eqs. (4.2) and (4.3) to be valid, one must also require the integrability condition $(J'_y)_z = (J'_z)_y$. If L represents any one of the differential operators in Eq. (2.6) [which operators appear on the right-hand side of Eq. (4.2)], this condition can be shown to require $[H_y, H_z] = 0$, where $H = (LJ)J^{-1}$. Thus this is a condition on the original solution J. Many solutions satisfy this condition, so that it is not excessively restrictive. This condition will be explored in future publications.

V. CONCLUSIONS AND SUMMARY

Let us summarize our main conclusions.

(1) The study of first-order Lie–Bäcklund type symmetries of the SDYM system requires the construction of a prolonged ideal of four-forms, which is many times larger than the ideal used for the derivation of point transformations. A noteworthy feature of the expanded ideal is the presence of the integrability conditions of the system.

(2) Starting with the (point) symmetries of SDYM on the base manifold, one can construct nine evolutionary vector fields that are generalized isovectors of the system. Nine nonlocal symmetries of SDYM are thus obtained. No further generalized symmetries can be generated by simply taking the Lie brackets of the nine EVF's.

(3) It is our conclusion that all evolutionary representatives of the corresponding point symmetries of SDYM yield Bäcklund transformations for the system. These BT's are "weak," in the sense that they relate two functions, of which the second is a solution of SDYM provided that the first one is. The physical implications of these transformations are not yet totally clear to us.

APPENDIX A: PROOF OF ISOVECTOR PROPERTY

We display the expansions of the Lie derivatives of the forms γ_1, γ_2, γ_3, with respect to the prolonged EVF's $\bar V_s$ ($s = 1,...,9$) of Eq. (3.16). We will use the notation

$$\gamma_i^{(s)} \equiv \underset{\bar V_s}{\pounds}\, \gamma_i \quad (i = 1,2,3).$$

Analytically,

$$\begin{aligned}
\gamma_1^{(1)} &= \gamma_5,\\
\gamma_2^{(1)} &= \gamma_2 B^1 + J\gamma_{10},\\
\gamma_3^{(1)} &= \gamma_3 B^1 - J\gamma_4 + J\gamma_{13},\\
\gamma_1^{(2)} &= \gamma_6 + \gamma_9 + [\gamma_{11}, B^2] + [B^1, \gamma_{15}] - \gamma_{20},\\
\gamma_2^{(2)} &= \gamma_2 B^2 + J\gamma_{13},\\
\gamma_3^{(2)} &= \gamma_3 B^2 + J\gamma_{14},\\
\gamma_1^{(3)} &= \gamma_7,\\
\gamma_2^{(3)} &= J\gamma_{11} + \gamma_{22} B^1 + \gamma_{25},\\
\gamma_3^{(3)} &= J\gamma_{15} + \gamma_{22} B^2 + \gamma_{27},\\
\gamma_1^{(4)} &= \gamma_{18},\\
\gamma_2^{(4)} &= J\gamma_{12} + \gamma_{23} B^1 + \gamma_{26},\\
\gamma_3^{(4)} &= J\gamma_1 - J\gamma_{11} + \gamma_{23} B^2 + \gamma_{28},\\
\gamma_1^{(5)} &= y\gamma_5 + \gamma_{11} + \bar z\gamma_{18},\\
\gamma_2^{(5)} &= \gamma_2 yB^1 + yJ\gamma_{10} + \bar zJ\gamma_{12} + \gamma_{23}\bar zB^1 + \bar z\gamma_{26},\\
\gamma_3^{(5)} &= \bar zJ\gamma_1 + \gamma_3 yB^1 - yJ\gamma_4 - \bar zJ\gamma_{11}\\
&\quad + yJ\gamma_{13} + \gamma_{23}\bar zB^2 + \bar z\gamma_{28},\\
\gamma_1^{(6)} &= \gamma_1 + z\gamma_6 + \bar y\gamma_7 + z\gamma_9 - \gamma_{11}\\
&\quad + [\gamma_{11}, zB^2] + [zB^1, \gamma_{15}] - z\gamma_{20},\\
\gamma_2^{(6)} &= \gamma_2 zB^2 + \bar yJ\gamma_{11} + zJ\gamma_{13} + \gamma_{22}\bar yB^1 + \bar y\gamma_{25},\\
\gamma_3^{(6)} &= \gamma_3 zB^2 + zJ\gamma_{14} + \bar yJ\gamma_{15} + \gamma_{22}\bar yB^2 + \bar y\gamma_{27},\\
\gamma_1^{(7)} &= z\gamma_5 + \gamma_{12} - \bar y\gamma_{18},\\
\gamma_2^{(7)} &= \gamma_2 zB^1 + zJ\gamma_{10} - \bar yJ\gamma_{12} - \gamma_{23}\bar yB^1 - \bar y\gamma_{26},\\
\gamma_3^{(7)} &= -\bar yJ\gamma_1 + \gamma_3 zB^1 - zJ\gamma_4 + \bar yJ\gamma_{11}\\
&\quad + zJ\gamma_{13} - \gamma_{23}\bar yB^2 - \bar y\gamma_{28},\\
\gamma_1^{(8)} &= y\gamma_6 - \bar z\gamma_7 + y\gamma_9 + [\gamma_{11}, yB^2]\\
&\quad + [yB^1, \gamma_{15}] + \gamma_{15} - y\gamma_{20},\\
\gamma_2^{(8)} &= \gamma_2 yB^2 - \bar zJ\gamma_{11} + yJ\gamma_{13} - \gamma_{22}\bar zB^1 - \bar z\gamma_{25},\\
\gamma_3^{(8)} &= \gamma_3 yB^2 + yJ\gamma_{14} - \bar zJ\gamma_{15} - \gamma_{22}\bar zB^2 - \bar z\gamma_{27},\\
\gamma_1^{(9)} &= \bar y\gamma_7 + \bar z\gamma_{18},\\
\gamma_2^{(9)} &= \bar yJ\gamma_{11} + \bar zJ\gamma_{12} + \gamma_{22}\bar yB^1\\
&\quad + \gamma_{23}\bar zB^1 + \bar y\gamma_{25} + \bar z\gamma_{26},\\
\gamma_3^{(9)} &= \bar zJ\gamma_1 - \bar zJ\gamma_{11} + \bar yJ\gamma_{15} + \gamma_{22}\bar yB^2\\
&\quad + \gamma_{23}\bar zB^2 + \bar y\gamma_{27} + \bar z\gamma_{28}.
\end{aligned}$$

APPENDIX B: PROOF OF BÄCKLUND TRANSFORMATIONS

Consider the nine pairs of parametric equations defined by Eqs. (4.2) and (4.3) (each pair share a common subscript k in δ_k). We show that each of these pairs is a (finite) BT for SDYM. For this purpose we cross-differentiate with respect to $\bar y$ and $\bar z$, and then add by terms, assuming that all integrability conditions are satisfied. We will use the notation

$$F[J] \equiv \partial_{\bar y}(J^{-1}J_y) + \partial_{\bar z}(J^{-1}J_z),$$

so that $F[J] = 0$ implies the SDYM equation. From the nine pairs of equations we thus obtain, respectively,

$$\begin{aligned}
F[J'] - F[J] &= \lambda\, \partial_y F[J],\\
F[J'] - F[J] &= \lambda\, \partial_z F[J],\\
F[J'] - F[J] &= \lambda\, \partial_{\bar y} F[J],\\
F[J'] - F[J] &= 0,\\
F[J'] - F[J] &= \lambda y\, \partial_y F[J],
\end{aligned}$$

$$F[J'] - F[J] = \lambda(1 + z\,\partial_z + \bar{y}\,\partial_{\bar{y}})F[J],$$
$$F[J'] - F[J] = \lambda z\,\partial_y F[J],$$
$$F[J'] - F[J] = \lambda(y\,\partial_z - \bar{z}\,\partial_{\bar{y}})F[J],$$
$$F[J'] - F[J] = \lambda\bar{y}\,\partial_{\bar{y}} F[J].$$

Thus $F[J] = 0$ implies $F[J'] = 0$ in each case, which establishes the Bäcklund transformation property.

[1] C. J. Papachristou and B. K. Harrison, J. Math. Phys. **28**, 1261 (1987).

[2] C. J. Papachristou and B. K. Harrison, in *Proceedings of the XV International Colloquium on Group Theoretical Methods in Physics*, edited by R. Gilmore (World Scientific, Singapore, 1987).

[3] B. K. Harrison and F. B. Estabrook, J. Math. Phys. **12**, 653 (1971); D. G. B. Edelen, *Applied Exterior Calculus* (Wiley, New York, 1985), Chap. 6.

[4] C. N. Yang, Phys. Rev. Lett. **38**, 1377 (1977); Y. Brihaye, D. B. Fairlie, J. Nuyts, and R. G. Yates, J. Math. Phys. **19**, 2528 (1978).

[5] M. K. Prasad, A. Sinha, and L.-L. Chau Wang, Phys. Rev. Lett. **43**, 750 (1979).

[6] L.-L. Chau and F. J. Chinea, Lett. Math. Phys. **12**, 189 (1986).

[7] P. J. Olver, *Applications of Lie Groups to Differential Equations, Graduate Texts in Mathematics* (Springer, New York, 1986), especially Chap. 5.

[8] We emphasize that the term "prolongation" is used in a very liberal sense in this article, and it may apply to variables, vector fields, differential equations, or ideals. Its precise meaning will always be clear from the context.

[9] A. S. Fokas, J. Math. Phys. **21**, 1318 (1980).

[10] H. C. Morris, J. Math. Phys. **21**, 256 (1980).

[11] This type of generalization can be supported by rigorous, general arguments that are not presented here. See, for example, P. H. M. Kersten, Ph.D. thesis, Twente University of Technology, The Netherlands, 1985.

[12] We write $B^1 = J^{-1}D_y J$ and $B^2 = J^{-1}D_z J$, where the x^μ and J are treated as independent variables at this stage (whence the use of total derivatives). Here, $D_y J$ and $D_z J$ are names for prolongation variables in the jet space, which reduce, upon projection to the solution manifold, simply to J_y and J_z.

[13] More generally, if $V = \xi^\mu\,\partial/\partial x^\mu + A\,\partial/\partial J$, then the evolutionary representative (3.14) of V is constructed by taking $Q = A - \xi^\mu D_\mu J$.

Transformation group acting on a self-dual Yang–Mills hierarchy

Yoshimasa Nakamura
Department of Mathematics, Faculty of Education, Gifu University, Gifu 501-11, Japan

(Received 16 April 1987; accepted for publication 2 September 1987)

A GL(n,C) self-dual Yang–Mills hierarchy is introduced; it is an infinite system of self-dual Yang–Mills equations having an infinite number of independent variables. Cauchy problems for the hierarchy are formally solved by using Lie transforms of a wave matrix. A relationship between the Kadomtsev–Petviashvili hierarchy and the self-dual Yang–Mills hierarchy is discussed. Furthermore, it is shown that an infinite-dimensional transformation group acts on a solution space to the ($n \geqslant 2$) self-dual Yang–Mills hierarchy. A parametric solution to the hierarchy is also given as a representation of the transformation group.

I. INTRODUCTION

The purpose of this paper is twofold. First, we introduce an infinite system, having an infinite number of variables, of self-dual Yang–Mills equations and call it a GL(n,C) self-dual Yang–Mills hierarchy after the Kadomtsev–Petviashvili (KP) hierarchy[1–5] and the Toda lattice hierarchy[6] (a discrete version of the KP hierarchy). Indeed, Sato and Sato[1] introduced the KP hierarchy and completely characterized the solution space to the KP equation as being of great interest. Here the KP hierarchy is an infinite system of integrable nonlinear evolution equations having an infinite number of time variables. Sato and Sato[1] also showed that many other integrable nonlinear equations called soliton equations are derived from the KP hierarchy by a reduction procedure; an infinite order pseudodifferential operator of the exponential type plays an essential role in their theory. Inspired by this pioneer work, Takasaki[7] proposed a method of generating formal power series solutions to the GL(n,C) self-dual Yang–Mills equation which seems to be *outside* of the KP hierarchy. An exponential operator also emerged in the representation of solutions. Further developments can be found in the works of Corrigan *et al.*[8] and Ward,[9] who discussed several sequences of self-dual Yang–Mills-like gauge field equations in dimensions higher than 4. However, it has not been clear how to define a self-dual Yang–Mills hierarchy. In the subsequent discussions, we shall treat this open problem and introduce an infinite system, in which every GL(kn,C) self-dual Yang–Mills equation ($k \in$N) is embedded, as a hopeful candidate for a self-dual Yang–Mills hierarchy.

The second purpose is to consider an infinite-dimensional transformation group acting on a solution space to the self-dual Yang–Mills hierarchy. It is also shown that a part of the infinitesimal actions of this group is very similar to the Kinnersley–Chitre (KC) transformation[10] of the Geroch group[11] in general relativity. Here the Geroch group is an infinite-dimensional transformation group acting on the stationary Einstein equations. Many classes of stationary space-times can be constructed by actions of the Geroch group.[12,13] Here we obtain a parametric solution to the self-dual Yang–Mills hierarchy as a representation of the transformation group.

In Sec. II, we first consider a system of linear algebraic and differential equations for an infinite matrix function. We then explain how a self-dual Yang–Mills hierarchy is characterized by this linear system. It is also shown that two types of Cauchy problems are formally solved by using solutions to the linear system. A relationship between the KP hierarchy and the self-dual Yang–Mills hierarchy is also discussed. In Sec. III, having defined a set of one-parameter transformations for the linear system, we construct a parametric solution to the self-dual Yang–Mills hierarchy. It is shown that the solution gives an infinite matrix representation of a transformation group acting on a solution space of the hierarchy. Section IV is devoted to concluding remarks. We shall resort to a method based on the linear algebra of infinite matrices throughout the following sections.

II. DEFINITION OF SELF-DUAL YANG–MILLS HIERARCHY

Let $W = W(y,\bar{y},z,\bar{z})$ be an $\infty \times \infty$ matrix function of the independent complex variables

$$\begin{aligned} y &= (y_1,y_2,\ldots), \quad \bar{y} = (\bar{y}_1,\bar{y}_2,\ldots), \\ z &= (z_1,z_2,\ldots), \quad \bar{z} = (\bar{z}_1,\bar{z}_2,\ldots), \end{aligned} \tag{2.1}$$

whose elements are arrayed as

$$W = (w_{i,j})_{i,j\in\mathbb{Z}} = \begin{bmatrix} & & \cdots & & \\ \cdots & w_{-1,-1} & w_{-1,0} & w_{-1,1} & \cdots \\ \cdots & w_{0,-1} & w_{0,0} & w_{0,1} & \cdots \\ \cdots & w_{1,-1} & w_{1,0} & w_{1,1} & \cdots \\ & & \cdots & & \end{bmatrix}. \tag{2.2}$$

We also use a semi-infinite matrix representation

$$W = \begin{bmatrix} W_{--} & W_{-+} \\ W_{+-} & W_{++} \end{bmatrix}, \tag{2.3}$$

where

$$W_{--} = (w_{-i,-j})_{i,j\in\mathbb{N}}, \quad W_{-+} = (w_{-i,j-1})_{i,j\in\mathbb{N}},$$
$$W_{+-} = (w_{i-1,-j})_{i,j\in\mathbb{N}}, \quad W_{++} = (w_{i-1,j-1})_{i,j\in\mathbb{N}}.$$

We call W_{--} a semi-infinite matrix of the $(--)$ type and so on. The product of the $\infty \times \infty$ matrices is defined by

$$(a_{i,j})_{i,j\in\mathbb{Z}} \cdot (b_{i,j})_{i,j\in\mathbb{Z}} = \Big(\sum_{k\in\mathbb{Z}} a_{i,k} b_{k,j}\Big)_{i,j\in\mathbb{Z}}.$$

Providing W_{++} is invertible, we introduce an infinite matrix function $\Xi = \Xi(y,\bar{y},z,\bar{z}) = (\xi_{i,j})_{i,j\in\mathbb{Z}}$ as

0022-2488/88/010244-05$02.50

$$\Xi=\begin{bmatrix}0_{--} & W_{-+}W^{-1}_{++}\\ 0_{+-} & 1_{++}\end{bmatrix}. \tag{2.4}$$

Here $1_{++}=(\delta_{i-1,j-1})_{i,j\in\mathrm{N}}$, where $\delta_{i-1,j-1}$ is the Kronecker delta and 0_{--} is the zero matrix of the $(--)$ type, etc.

Let n be a positive integer and Λ be a shift matrix defined by

$$\Lambda=(\delta_{1-i,-j})_{i,j\in\mathrm{Z}}. \tag{2.5}$$

Here we impose a linear algebraic constraint on W:

$$[\Lambda^n,W]=0, \tag{2.6}$$

where $\Lambda^n=(\delta_{n-i,j})_{i,j\in\mathrm{Z}}$. This implies that W is a (block) Toeplitz matrix. It is easy to see $[\Lambda^{kn},W]=0$ for any $k\in\mathrm{N}$. We now have the following useful identities (a generalization of the identity found in Ref. 7).

Lemma 1: For any $k\in\mathrm{N}$, we have

$$\Lambda^{kn}\Xi=\Xi\Lambda^{kn}\Xi. \tag{2.7}$$

Proof: We set

$$\Lambda^{kn}=\begin{bmatrix}\Lambda^{kn}_{--} & 0_{-+}\\ \Lambda^{kn}_{+-} & \Lambda^{kn}_{++}\end{bmatrix}.$$

Using $[\Lambda^{kn},W]=0$, we have

$$\begin{aligned}\Lambda^{kn}_{--}W_{-+}W^{-1}_{++}&=W_{-+}\Lambda^{kn}_{++}W^{-1}_{++}\\ &=W_{-+}W^{-1}_{++}(\Lambda^{kn}_{+-}W_{-+}\\ &\quad+\Lambda^{kn}_{++}W_{++})W^{-1}_{++}\\ &=W_{-+}W^{-1}_{++}\Lambda^{kn}_{+-}W_{-+}W^{-1}_{++}\\ &\quad+W_{-+}W^{-1}_{++}\Lambda^{kn}_{++},\end{aligned}$$

for $k\in\mathrm{N}$. This immediately gives (2.7). □

Next we consider the second constraint on W. Let W be a solution to the infinite system of compatible linear differential equations:

$$\begin{aligned}&D_kW(y,\bar y,z,\bar z)=0,\quad D_k=\partial_{y_k}+\Lambda^{kn}\partial_{\bar z_k},\\ &D^*_lW(y,\bar y,z,\bar z)=0,\quad D^*_l=\partial_{z_l}-\Lambda^{ln}\partial_{\bar y_l}\end{aligned} \tag{2.8}$$

for $k,l\in\mathrm{N}$, where $\partial_{y_k}=\partial/\partial_{y_k}$, etc. Since W is Toeplitz, we can write $W=(W_{j-i})_{i,j\in\mathrm{Z}}$, where W_{j-i} are $n\times n$ matrices. Observe that (2.8) imply $\partial_{y_k}W_{j-i}+\partial_{\bar z_k}W_{j-i+k}=0$ and $\partial_{z_l}W_{j-i}-\partial_{\bar y_l}W_{j-i+l}=0$. We see that the W_{j-i} satisfy the matrix Laplace equations $(\partial_{y_k}\partial_{\bar y_k}+\partial_{z_k}\partial_{\bar z_k})W_{j-i}=0$ providing the W_{j-i} are integrable. In the following discussions we assume the integrability of W w.r.t. $(y,\bar y,z,\bar z)$. Making use of the identities (2.7) as well as (2.6), we have

$$\begin{aligned}\partial_{y_k}(W_{-+}W^{-1}_{++})&=(-W_{-+}W^{-1}_{++}\partial_{y_k}W_{++}+\partial_{y_k}W_{-+})W^{-1}_{++}\\ &=\{W_{-+}W^{-1}_{++}(\Lambda^{kn}_{+-}\partial_{\bar z_k}W_{-+}+\Lambda^{kn}_{++}\partial_{\bar z_k}W_{++})-\Lambda^{kn}_{--}\partial_{\bar z_k}W_{-+}\}W^{-1}_{++}\\ &=W_{-+}W^{-1}_{++}\Lambda^{kn}_{+-}\partial_{\bar z_k}(W_{-+}W^{-1}_{++})-\Lambda^{kn}_{--}\partial_{\bar z_k}(W_{-+}W^{-1}_{++})+(W_{-+}W^{-1}_{++}\Lambda^{kn}_{++}\\ &\quad+W_{-+}W^{-1}_{++}\Lambda^{kn}_{+-}W_{-+}W^{-1}_{++}-\Lambda^{kn}_{--}W_{-+}W^{-1}_{++})\partial_{\bar z_k}W_{++}\cdot W^{-1}_{++}\\ &=W_{-+}W^{-1}_{++}\Lambda^{kn}_{+-}\partial_{\bar z_k}(W_{-+}W^{-1}_{++})-\Lambda^{kn}_{--}\partial_{\bar z_k}(W_{-+}W^{-1}_{++})\end{aligned}$$

from $D_kW=0$. Since $W_{-+}W^{-1}_{++}$ is the $(-\;+)$ part of Ξ [see (2.4)], the above equations are expressed as

$$D_k\Xi=\Xi\Lambda^{kn}\partial_{\bar z_k}\Xi, \tag{2.9a}$$

for $k\in\mathrm{N}$. Similarly, from $D^*_lW=0$ in (2.8) we derive

$$D^*_l\Xi=-\Xi\Lambda^{ln}\partial_{\bar y_l}\Xi, \tag{2.9b}$$

for $l\in\mathrm{N}$. Observe that if Ξ satisfies

$$\begin{aligned}&D_k(\Lambda^{ln}\partial_{\bar z_l}\Xi)+D^*_l(\Lambda^{kn}\partial_{\bar y_k}\Xi)\\ &\quad+[\Lambda^{ln}\partial_{\bar z_l}\Xi,\Lambda^{kn}\partial_{\bar y_k}\Xi]=0,\end{aligned}$$

then $[D_k,D^*_l]\Xi=0$ from (2.9a) and (2.9b). We derive $\partial_{\bar y_k}\partial_{\bar z_l}\Xi=\partial_{\bar z_l}\partial_{\bar y_k}\Xi$ from $\partial_{\bar y_k}\partial_{\bar z_l}W=\partial_{\bar z_l}\partial_{\bar y_k}W$. Hence if Ξ satisfies the second-order nonlinear equations

$$\begin{aligned}&\partial_{z_k}(\Lambda^{ln}\partial_{\bar z_l}\Xi)+\partial_{y_l}(\Lambda^{kn}\partial_{\bar y_k}\Xi)\\ &\quad+[\Lambda^{ln}\partial_{\bar z_l}\Xi,\Lambda^{kn}\partial_{\bar y_k}\Xi]=0,\end{aligned} \tag{2.10}$$

for $k,l\in\mathrm{N}$, then (2.9a) and (2.9b) are mutually compatible. The remaining compatibility conditions are

$$\begin{aligned}&\partial_{y_l}(\Lambda^{kn}\partial_{\bar z_k}\Xi)-\partial_{y_k}(\Lambda^{ln}\partial_{\bar z_l}\Xi)\\ &\quad-[\Lambda^{kn}\partial_{\bar z_k}\Xi,\Lambda^{ln}\partial_{\bar z_l}\Xi]=0,\\ &\partial_{z_l}(\Lambda^{kn}\partial_{\bar y_k}\Xi)-\partial_{z_k}(\Lambda^{ln}\partial_{\bar y_l}\Xi)\\ &\quad+[\Lambda^{kn}\partial_{\bar y_k}\Xi,\Lambda^{ln}\partial_{\bar y_l}\Xi]=0,\end{aligned} \tag{2.11}$$

for $k,l\in\mathrm{N}$. The system (2.10) and (2.11) has an infinite number of independent variables $(y,\bar y,z,\bar z)$.

Every GL(kn,C) self-dual Yang–Mills equation $(k\in\mathrm{N})$ is embedded in (2.10). In order to see this, let us define the $ln\times ln$ matrices

$$\Xi^{(i,j)}_l=\begin{bmatrix}\xi_{-(i+1)ln,jln} & \cdots & \xi_{-(i+1)ln,(j+1)ln-1}\\ \vdots & & \vdots\\ \xi_{-iln-1,jln} & \cdots & \xi_{-iln-1,(j+1)ln-1}\end{bmatrix} \tag{2.12}$$

for $i,j\in\mathrm{N}\cup\{0\}$, $l\in\mathrm{N}$. It is shown from (2.10) that

$$\begin{aligned}&\partial_{z_k}\partial_{\bar z_k}\Xi^{(0,0)}_k+\partial_{y_k}\partial_{\bar y_k}\Xi^{(0,0)}_k\\ &\quad+[\partial_{\bar z_k}\Xi^{(0,0)}_k,\partial_{\bar y_k}\Xi^{(0,0)}_k]=0,\end{aligned} \tag{2.13}$$

for $k\in\mathrm{N}$. Since (2.13) are zero-curvature conditions, there are GL(kn,C)-valued functions $G_k=G_k(y,\bar y,z,\bar z)$ such that

$$\partial_{\bar z_k}\Xi^{(0,0)}_k=\partial_{y_k}G_k\cdot G^{-1}_k,\quad \partial_{\bar y_k}\Xi^{(0,0)}_k=-\partial_{z_k}G_k\cdot G^{-1}_k, \tag{2.14}$$

for $k\in\mathrm{N}$. Hence each G_k satisfies the GL(kn,C) self-dual Yang–Mills equation,

$$\partial_{\bar y_k}(\partial_{y_k}G_k\cdot G^{-1}_k)+\partial_{\bar z_k}(\partial_{z_k}G_k\cdot G^{-1}_k)=0, \tag{2.15}$$

for $k\in\mathrm{N}$. The usual GL(n,C) equation[7,14] on a complexified four-dimensional space is derived from (2.15) by setting $k=1$. The remaining part of (2.10) determines a sequence of potentials for (2.15). We shall call the infinite system (2.10) and (2.11) with the constraint (2.7) *a* $\mathrm{GL}(n,\mathrm{C})$ *self-dual Yang–Mills hierarchy* after the KP hierarchy[1–5] and the Toda lattice hierarchy.[6] Conversely, the system (2.8) with (2.6) is said to be a linearization of the self-dual Yang–Mills hierarchy. We also call a solution to (2.6) and (2.8) a wave matrix. By assumption, w_{ij} satisfy four-dimensional Laplace equations and hence solutions to the self-dual Yang–Mills hierarchy can be given by nonlinear superpositions of harmonic functions through the use of (2.4).

We remark that Eqs. (2.9a) and (2.9b) define an infinite number of commutative flows on an infinite-dimensional Grassman manifold whose affine coordinates are given by $W_{-+}W_{++}^{-1}$. This geometric interpretation will be justified by a line of thought similar to that given in Refs. 1, 7, and 15.

Next we consider Cauchy problems for the self-dual Yang–Mills hierarchy with the help of linearization discussed above. For a given initial data $W^{\text{in}}=W(0,\bar{y},0,\bar{z})$, the compatible linear system (2.8) is solved locally (in y and z) as

$$W(y,\bar{y},z,\bar{z})=\exp\left\{\sum_{l\in\mathrm{N}}(z_l\Lambda^{ln}\partial_{jy_l}-y_l\Lambda^{ln}\partial_{\bar{z}_l})\right\}W^{\text{in}},\qquad(2.16)$$

where

$$\exp(X)=\sum_{k\in\mathrm{N}\cup\{0\}}\frac{1}{k!}X^k.$$

The solution (2.16) is a Lie transform of W^{in}. If W^{in} is integrable w.r.t. $(\bar{y}_k,\bar{z}_l)$, then $W=W(y,\bar{y},z,\bar{z})$ is as well. If W^{in} satisfies (2.6), then W also does so. Furthermore, if W^{in}_{++}, the $(++)$ part of W^{in}, is invertible, then W_{++} is as well. Hence (2.16) describes a "time evolution" of the self-dual Yang–Mills hierarchy. We have proved a generalization of Takasaki's reconstruction formula for the single self-dual Yang–Mills equation.[7]

Proposition 2: Let

$$\Xi^{\text{in}}=\begin{bmatrix}0_{--} & W^{\text{in}}_{-+}W^{\text{in}-1}_{++}\\ 0_{+-} & 1_{++}\end{bmatrix}$$

be an initial data for (2.10) such that $\partial_{\bar{y}_k}\partial_{\bar{z}_l}W^{\text{in}}=\partial_{\bar{z}_l}\partial_{\bar{y}_k}W^{\text{in}}$ for $k,l\in\mathrm{N}$ and $[\Lambda^n,W^{\text{in}}]=0$. Then a formal power series

$$\Xi=\begin{bmatrix}0_{--} & W_{-+}W^{-1}_{++}\\ 0_{+-} & 1_{++}\end{bmatrix}$$

satisfies the self-dual Yang–Mills hierarchy (2.10) and (2.11) with (2.7).

It should be noted that there are two time evolution directions for the self-dual Yang–Mills hierarchy. Let $W^{\text{in}}=W(y,0,z,0)$ be an initial data to (2.8). Then a Lie transform

$$W(y,\bar{y},z,\bar{z})=\exp\left\{\sum_{l\in\mathrm{N}}(\bar{y}_l\Lambda^{-ln}\partial_{z_l}-jz_l\Lambda^{-ln}\partial_{y_l})\right\}W^{\text{in}}\qquad(2.17)$$

of W^{in} gives another formal power series solution to the self-dual Yang–Mills hierarchy provided that $\partial_{y_k}\partial_{z_l}W^{\text{in}}=\partial_{z_l}\partial_{y_k}W^{\text{in}}$ for $k,l\in\mathrm{N}$ and $[\Lambda^n,W^{\text{in}}]=0$. The proof is carried out in the same manner as in Proposition 2. Formula (2.17) describes the second time evolution.

In what follows of this section we trace the relationship between the KP hierarchy and the self-dual Yang–Mills hierarchy. Let us consider one-parameter transformations for the wave matrix defined by

$$(\partial_{s_m}-\Lambda^m)W(y,\bar{y},z,\bar{z},s)=0,\qquad(2.18)$$

for $m\in\mathrm{N}$, where $s=(s_1,s_2,\ldots)$ is a set of complex variables. If $W(y,\bar{y},z,\bar{z};0)$ is a wave matrix, then the Lie transform

$$W(y,\bar{y},z,\bar{z},s)=\exp\left(\sum_{m\in\mathrm{N}}s_m\Lambda^m\right)W(y,\bar{y},z,\bar{z},0)\qquad(2.19)$$

is essentially equivalent to the time evolution for the Toda lattice hierarchy[6] (a discrete version of the KP hierarchy). Hence we can associate the time variables of the KP hierarchy with the variables s. The corresponding matrix $\Xi=\Xi(y,\bar{y},z,\bar{z},s)$ satisfies the compatible nonlinear equations

$$\partial_{s_m}\Xi=\Lambda^m\Xi-\Xi\Lambda^m\Xi,\qquad(2.20)$$

for $m\in\mathrm{N}$. We then see from the identities (2.7) that $\partial_{s_m}\Xi=0$ for $m=0\ (\mathrm{mod}\ n)$, that is, Ξ does not depend on s_m for $m=0\ (\mathrm{mod}\ n)$. The condition (2.6) can be regarded as an n-periodic condition for the wave matrix W. Every nth time variable of the KP hierarchy is related to hidden variables. The other time variables survive in Ξ. This fact makes clear an important connection between the KP hierarchy and our self-dual Yang–Mills hierarchy.

Finally, we refer to the recent works on formal power solutions to the Witten equation[16] (a higher-dimensional generalization of self-dual Yang–Mills equation), the supersymmetric Yang–Mills equation,[17] and the stationary axially symmetric Einstein equations.[18,19]

III. TRANSFORMATION GROUP

Let $t=(\ldots,t_{-1},t_0,t_1,\ldots)$ be a set of complex parameters and h_k, $k\in\mathbf{Z}$, be $\mathrm{gl}(n,\mathrm{C})$-valued constant matrices. Define

$$H_k=\mathrm{diag}(h_k)\Lambda^{kn},\qquad(3.1)$$

for $k\in\mathbf{Z}$, where $\mathrm{diag}(h_k)$ denotes an infinite (block) diagonal matrix. We note that $[\Lambda^{ln},H_k]=0$ for $k,l\in\mathbf{Z}$. Let us consider a one-parameter transformation for the wave matrix W defined by

$$(\partial_{t_k}-H_k)W(y,\bar{y},z,\bar{z};t)=0,\qquad(3.2)$$

for some $k\in\mathbf{Z}$. A parametric solution $W(t)=W(y,\bar{y},z,\bar{z};t)$ to the system (3.2) gives $\Xi(t)=\Xi(y,\bar{y},z,\bar{z};t)$ by (2.4), which satisfies

$$\begin{aligned}\partial_{t_{-l}}\Xi(t)&=H_{-l}\Xi(t)-\Xi(t)H_{-l},\\ \partial_{t_k}\Xi(t)&=H_k\Xi(t)-\Xi(t)H_k\Xi(t),\end{aligned}\qquad(3.3)$$

for $l\in\mathrm{N}\cup\{0\}$ and $k\in\mathrm{N}$. The proof of (3.3) is carried out in a manner similar to (2.9a). If we focus on the $n\times n$ matrices $\Xi_1^{(i,j)}$ defined by (2.12), we have, from the second term in (3.3),

$$\partial_{t_k}\Xi_1^{(i,j)} = h_k\Xi_1^{(i+k,j)} - \Xi_1^{(i,j+k)}h_k - \sum_{l=0}^{k-1}\Xi_1^{(i,l)}h_k\Xi_1^{(k-l-1,j)}, \qquad (3.4)$$

for $i, j \in \mathrm{N}\cup\{0\}$, $k\in\mathrm{N}$. The transformation described by (3.4) is very similar to the KC transformation[10] of the Geroch group[11] acting on the stationary Einstein equations. Analogs of the KC transformation for the self-dual Yang–Mills equation have been known as an Ehlers-type transformation[20] and an infinitesimal Riemann–Hilbert transformation[14]; however, a group theoretic structure of the totality has been obscure. We note that the general $kn\times kn$ matrices $\Xi_k^{(i,j)}$ satisfy

$$\partial_{t_k}\Xi_k^{(i,j)} = \mathrm{diag}[h_k]\Xi_k^{(i+1,j)} - \Xi_k^{(i,j+1)}\mathrm{diag}[h_k] - \Xi_k^{(i,0)}\mathrm{diag}[h_k]\Xi_k^{(0,j)} \qquad (3.5)$$

for $i,j\in\mathrm{N}\cup\{0\}$ and $k\in\mathrm{N}$, where $\mathrm{diag}[h_k]$ denotes the $kn\times kn$ (block) diagonal matrix.

We now exponentiate the linear system (3.2) for a given data which satisfies the linear system (2.8) with (2.6). We can show after a calculation the following.

Lemma 3: Let $W^0 = W(y,\bar{y},z,\bar{z};0)$, satisfying (2.6) and (2.8), be an initial data for (3.2). Then a solution

$$W(t_k) = \exp(t_kH_k)W^0 \qquad (3.6)$$

to (3.2) for some $k\in\mathrm{N}$ also satisfies (2.6) and (2.8).

Lemma 3 implies that $\Xi(t_k)$ corresponding to $W(t_k)$ gives a solution to the self-dual Yang–Mills hierarchy. In other words, the one-parameter transformation (3.6) for W gives rise to a symmetry of the hierarchy. It is noted that if W^0_{++} is invertible then $W_{++}(t)$ is as well. Here we have a useful representation of $\Xi(t_k)$ as follows.

Proposition 4: Let

$$\Xi^0 = \begin{bmatrix} 0_{--} & W^0_{-+}\mathrm{W}^{0-1}_{++} \\ 0_{+-} & 1_{++} \end{bmatrix}$$

be an initial data for (3.3) such that W^0 satisfies (2.6) and (2.8). Then $\Xi(t_k)$ defined by

$$\Xi(t_k) = \exp(t_kH_k)\Xi^0\cdot\{I - \Xi^0 + \exp(t_kH_k)\Xi^0\}^{-1}, \qquad (3.7)$$

for some $k\in\mathrm{Z}$, is a parametric solution to the self-dual Yang–Mills hierarchy (2.10) and (2.11) with (2.7), which also satisfies (3.3).

Proof: Set

$$\begin{bmatrix} \Gamma_{--} & 0_{-+} \\ \Gamma_{+-} & \Gamma_{++} \end{bmatrix} = \exp(t_kH_k),$$

for $k\in\mathrm{N}$. Then the rhs of (3.7) for $k\in\mathrm{N}$ is

$$\begin{bmatrix} 0_{--} & \Gamma_{--}W^0_{-+}W^{0-1}_{++} \\ 0_{+-} & \Gamma_{+-}W^0_{-+}W^{0-1}_{++} + \Gamma_{++} \end{bmatrix} \begin{bmatrix} 1_{--} & \Gamma_{--}W^0_{-+}W^{0-1}_{++} - W^0_{-+}W^{0-1}_{++} \\ 0_{+-} & \Gamma_{+-}W^0_{-+}W^{0-1}_{++} + \Gamma_{++} \end{bmatrix}^{-1}$$
$$= \begin{bmatrix} 0_{--} & \Gamma_{--}W^0_{-+}W^{0-1}_{++}(\Gamma_{+-}W^0_{-+}W^{0-1}_{++} + \Gamma_{++})^{-1} \\ 0_{+-} & 1_{++} \end{bmatrix} = \begin{bmatrix} 0_{--} & W_{-+}(t_k)W_{++}(t_k)^{-1} \\ 0_{+-} & 1_{++} \end{bmatrix} = \Xi(t_k).$$

The proof of (3.7) for $-k\in\mathrm{N}\cup\{0\}$ is carried out by setting

$$\begin{bmatrix} \Gamma_{--} & \Gamma_{-+} \\ 0_{+-} & \Gamma_{++} \end{bmatrix} = \exp(t_kH_k).$$

We omit it here. Taking account of Lemma 3, we see that $\Xi(t_k)$ is a solution to the self-dual Yang–Mills hierarchy. Furthermore, we can show that $\Xi(t_k)$ satisfies (3.3) by a direct calculation. This completes the proof. □

Next we discuss an algebraic structure of a whole set of one-parameter transformations induced by $\{H_k;\ k\in\mathrm{Z}\}$. We note that

$$[H_k,H_l] = \mathrm{diag}([h_k,h_l])\Lambda^{(k+l)n} \qquad (3.8)$$

for $k, l\in\mathrm{Z}$. Note that $[H_{k+m},H_{l-m}] = [H_k,H_l]$, providing $[h_{k+m},h_{l-m}] = [h_k,h_l]$. We then see that an algebra which the $\{H_k;\ k\in\mathrm{Z}\}$ action on the wave matrix forms is homomorphic to the infinite-dimensional graded Lie algebra, a subalgebra of the Lie algebra $\mathrm{gl}(\infty)$,

$$\mathrm{gl}(n,\mathrm{C})\otimes\mathrm{C}[\lambda,\lambda^{-1}], \qquad (3.9)$$

where λ is a complex parameter such that $|\lambda| = 1$. Hence it is concluded from Proposition 4 that the $\{H_k;\ k\in\mathrm{N}\}$ action on Ξ forms a Lie algebra homomorphic to (3.9). In mathematical literature, the algebra (3.9) is called the Kac–Moody algebra without center,[21] or the loop algebra.[22] Let us recall the fact that the KP hierarchy and its solutions are obtained from a representation of $\mathrm{gl}(\infty)$; subhierarchies such as the Korteweg–de Vries hierarchy and the Boussinesq hierarchy are associated with infinite-dimensional subalgebras of $\mathrm{gl}(\infty)$.[1,2] This suggests that our self-dual Yang–Mills hierarchy may be derived from a $\mathrm{GL}(\infty)$-invariant hierarchy by a reduction procedure.

The infinitesimal property discussed above reflects a group theoretical structure of the totality of transformations (3.7). We now come to the main theorem.

Theorem 5: The transformations (3.7) form a group $\mathbb{G}$ acting on a solution space to the self-dual Yang–Mills hierarchy.

Proof: We first note that a set of $\exp(t_kH_k)$, $k\in\mathrm{Z}$, forms a Lie group $\mathrm{GL}(\infty)$ w.r.t. the matrix multiplication. For H_k, H_l, and any initial data Ξ^0, let us set

$$\Xi(t_k,t_l) = \exp(t_lH_l)\Xi(t_k)\cdot\{I - \Xi(t_k) + \exp(t_lH_l)\Xi(t_k)\}^{-1}, \qquad (3.10)$$

where $\Xi(t_k)$ is as in (3.7). We then have a product formula of actions

$$\Xi(t_k,t_l) = \exp(t_lH_l)\exp(t_kH_k)\Xi^0 \times\{I - \Xi^0 + \exp(t_lH_l)\exp(t_kH_k)\Xi^0\}^{-1}. \qquad (3.11)$$

Equation (3.11) is checked by inserting (3.7) into the rhs of (3.10). The unit element of $\mathbb{G}$ is the identity action induced by $H_k = 0$. We also have a Baker–Campbell–Hausdorff formula

$$\begin{aligned}\exp(t_l H_l)\exp(t_k H_k) &= \exp\{t_k H_k + t_l H_l - (t_k t_l/2)[H_k,H_l] \\ &\quad + (t_k t_l^2/12)[H_l,[H_l,H_k]] \\ &\quad + (t_k^2 t_l/12)[H_k,[H_k,H_l]] + \cdots\}. \qquad (3.12)\end{aligned}$$

Hence the inverse action to any H_k is the action induced by $-H_k$. This completes the proof. □

Finally we remark on the following. Formula (3.12) shows that the infinite-dimensional transformation group having the representation (3.7) acts on the self-dual Yang–Mills hierarchy in the case where $n \geqslant 2$. If we choose $n = 1$ in (3.1), then $[H_k,H_l] = 0$ for any h_k and h_l. This implies that

$$\exp(t_l H_l)\exp(t_k H_k) = \exp(t_k H_k)\exp(t_l H_l)$$

from (3.12) and consequently, $\Xi(t_k,t_l) = \Xi(t_l,t_k)$ from (3.11). Hence the one-parameter transformation group having the representation (3.7) acts on the $n = 1$ self-dual Yang–Mills hierarchy (a Laplace equation hierarchy).

IV. CONCLUDING REMARKS

We have presented a GL(n,C) self-dual Yang–Mills hierarchy having an infinite number of independent variables. Cauchy problems for the hierarchy are formally solved by using Lie transforms. A relationship between the KP hierarchy and our self-dual Yang–Mills hierarchy is discussed (Sec. II). It is shown that an infinite-dimensional transformation group $\mathbb{G}$ acts on a solution space to the ($n \geqslant 2$) hierarchy. We do not feel that we are simply performing a formal theory. A parametric solution to the hierarchy is obtained explicitly as a representation of the transformation group (Sec. III). We see that our self-dual Yang–Mills hierarchy is integrable in the sense that we can integrate it by essentially linear techniques.

With these results comes a related problem that is being pursued. We wish to know if we can find a GL(∞)-invariant self-dual Yang–Mills hierarchy from which our GL(n,C) self-dual Yang–Mills hierarchy can be derived by a reduction procedure. If it is possible, the resulting GL(∞)-invariant hierarchy will be regarded as a natural generalization of the KP hierarchy to higher dimensions. We should remember that the transformation group for the KP hierarchy is the automorphism group GL(∞) of an infinite-dimensional Grassmann manifold and that many hierarchies of soliton equations are derived from the KP hierarchy by reductions of GL(∞).[1,2] We believe the self-dual Yang–Mills hierarchy introduced in this paper will play an important role in solving this interesting problem.

ACKNOWLEDGMENTS

The author would like to thank Dr. K. Takasaki, Dr. K. Ueno, Dr. K. Nagatomo, and Dr. T. Iwai for valuable discussions and helpful comments.

[1] M. Sato and Y. Sato, "Soliton equations as dynamical systems on infinite dimensional Grassman manifold," in *Nonlinear Partial Differential Equations in Applied Science; Proceedings of the U.S.–Japan Seminar, Tokyo, 1982*, edited by H. Fujita, P. D. Lax, and G. Strang (North-Holland, Amsterdam, 1983).

[2] M. Jimbo and T. Miwa, Publ. Res. Inst. Math. Sci. Kyoto Univ. **19**, 943(1983); E. Date, M. Jimbo, M. Kashiwara, and T. Miwa, "Transformation groups for soliton equations," in *Non-linear Integrable Systems–Classical Theory and Quantum Theory*, edited by M. Jimbo and T. Miwa (World Scientific, Singapore, 1983).

[3] M. Mulase, Adv. Math. **54**, 57 (1984).

[4] K. Ueno, "The Riemann-Hilbert decomposition and the KP hierarchy," in *Vertex Operators in Mathematics and Physics*, edited by J. Lepowsky, S. Mandelstam, and I. M. Singer (Springer, New York, 1985).

[5] B. Dorizzi, B. Grammaticos, A. Ramani, and P. Winternitz, J. Math. Phys. **27**, 2848 (1986).

[6] K. Ueno and K. Takasaki, "Toda lattice hierarchy," in *Group Representations and Systems of Differential Equations*, edited by K. Okamoto (Kinokuniya, Tokyo, 1984).

[7] K. Takasaki, Commun. Math. Phys. **94**, 35 (1984); Saitama Math. J. **3**, 11 (1985).

[8] E. Corrigan, C. Devchand, D. B. Fairlie, and J. Nuyts, Nucl. Phys. B **214**, 452 (1983).

[9] R. S. Ward, Nucl. Phys. B **236**, 381 (1984).

[10] W. Kinnersley and D. M. Chitre, J. Math. Phys. **19**, 2037 (1978).

[11] R. Geroch, J. Math. Phys. **13**, 394 (1972).

[12] C. Hoenselaers, W. Kinnersley, and B. C. Xanthopoulos, J. Math. Phys. **20**, 2530 (1979).

[13] Y. Nakamura, J. Math. Phys. **24**, 606 (1983).

[14] K. Ueno and Y. Nakamura, Phys. Lett. B **109**, 273 (1982); Publ. Res. Inst. Math. Sci. Kyoto Univ. **19**, 519 (1983).

[15] Y. Nakamura, "Riemann-Hilbert transformations for a Toeplitz matrix equation: Some ideas and applications to linear prediction problem," preprint, Gifu University.

[16] N. Suzuki, Proc. Jpn. Acad. A **60**, 141, 252 (1984).

[17] J. Harnad and M. Jacques, J. Math. Phys. **27**, 2394 (1986).

[18] Y. Nakamura, Class. Quant. Grav. **4**, 437 (1987).

[19] K. Nagatomo, "Formal power series solutions of the stationary axisymmetric vacuum Einstein equations," preprint, Osaka University.

[20] Y. Nakamura, Lett. Math. Phys. **7**, 171 (1983).

[21] V. G. Kac, *Infinite Dimensional Lie Algebras* (Birkhäuser, Boston, 1983).

[22] A. Pressley and G. Segal, *Loop Groups* (Oxford U.P., Oxford, 1986).

Normal forms of an abstract Dirac operator and applications to scattering theory

Bernd Thaller
Institut für Mathematik, University of Graz, A-8010 Graz, Austria

(Received 19 February 1987; accepted for publication 2 September 1987)

The unitary transformations which convert an abstract Dirac operator into an "even" (resp. "odd") operator are determined. The problem is formulated and solved completely within the general setup of supersymmetric quantum mechanics. This leads to some apparently new applications in relativistic quantum mechanics, where the transformations are known as the Foldy–Wouthuysen (resp. Cini–Touschek) transformations. The scattering theory for abstract Dirac operators is discussed and the utility of the general theory is illustrated by proving existence of relativistic MØller operators for scattering from long-range magnetic fields.

I. INTRODUCTION

Consider the complex 2×2 matrix

$$\begin{pmatrix} m & \bar{z} \\ z & -m \end{pmatrix}, \quad z\in\mathbb{C}, \quad m\in\mathbb{R}\,.$$

It is Hermitian and can therefore be diagonalized with the help of a unitary matrix. We want to investigate the self-adjoint operator H which is obtained from (1.1) if z (resp. $\bar{z}$) is replaced by a closed operator D in a Hilbert space (resp. D^*, the adjoint) and m by a symmetric operator commuting with D. More precisely, we consider

$$H=\begin{pmatrix} M_+ & D^* \\ D & -M_- \end{pmatrix}, \tag{1.1}$$

where $D^*M_- = M_+D^*$, $DM_+ = M_-D$ (see Sec. III for the precise definitions). A typical example for H is given by the Dirac operator in relativistic quantum mechanics. We shall therefore call H an abstract Dirac operator. In order to investigate the spectral and scattering theory for H we ask whether we can find—in analogy to the scalar case (1.1) above—a unitary transformation U bringing H to diagonal form.

For special operators D problems like this have been studied intensively in the context of relativistic quantum mechanics since the 1950's (see, e.g., Ref. 1 for a review). The main purpose was to find transformations bringing relativistic wave equations to "normal forms" which were considered to be "more canonical" than the original equations.[2] Most prominent became the so-called Foldy–Wouthuysen (FW) and Cini–Touschek (CT) transformations.[3,4] They have been obtained explicitly only in a number of special cases and sometimes only by formal methods.[1,5–7]

The abstract Dirac operator (1.1) contains all these above-mentioned special cases. Some examples covered by our approach are given in Sec. II. In Secs. III–V the problem of bringing the abstract Dirac operator to an "even" (resp. "odd") normal form is formulated rigorously and solved completely within the abstract framework of supersymmetric quantum mechanics. Supersymmetry plays an essential role in many fields of physics and mathematics.[8–12] In Refs. 8 and 9 applications to index theory and differential geometry are discussed. Examples in acoustics and optics may be extracted from Ref. 13.

In Sec. VI, the general results are applied to the scattering theory for the abstract Dirac operator. We prove existence of wave operators by reducing the problem to the study of an associated supersymmetric scattering system. As an example we discuss the relativistic scattering from long-range magnetic fields, which by our method is reduced to the corresponding nonrelativistic problem.

In many situations of physical interest the Dirac operator does not have the form (1.1) (e.g., for the Coulomb problem). In these cases an exact diagonal form has not been found so far. One has tried to obtain it by successive transformations which, if applied to the Dirac operator, leads to a series expansion consisting only of diagonal (even) operators.[3] We pay no attention to this because (1) as an operator series this expansion is divergent,[14,15] and (2) for the practical calculation of relativistic corrections to bound state energies it can be replaced by a mathematically rigorous method.[16,17]

II. SOME CONCRETE REALIZATIONS

In this section we give some examples of operators of the form (1.1). Note that this operator is naturally defined on the direct sum $\mathscr{H} = \mathscr{H}_+ \oplus \mathscr{H}_-$ of two Hilbert spaces $\mathscr{H}_+$, $\mathscr{H}_-$. All examples are taken from relativistic quantum mechanics. It is remarkable that also the Klein–Gordon equation, if suitably interpreted as a first-order system, can be written in Schrödinger form with an abstract Dirac operator as Hamiltonian (but with $\mathscr{H}_+ \neq \mathscr{H}_-$). The explicit form of the Foldy–Wouthuysen transformation U and the Cini–Touschek transformation V will be given in Secs. IV and V essentially in terms of D and $M_\pm$.

A. The free Dirac operator

The free time evolution of a spin-$\frac{1}{2}$ particle with mass m is generated by (c denotes the velocity of light)

$$H_0 = c\boldsymbol{\alpha}\cdot\mathbf{P} + \beta mc^2\,, \tag{2.1}$$

which is self-adjoint on the Sobolev space

$$\mathscr{D}(H_0) = W^{1,2}(\mathbf{R}^3)^4 \subset L^2(\mathbf{R}^3)^4 \equiv \mathscr{H}\,. \tag{2.2}$$

Here $\mathbf{P} = -i\nabla$, acting componentwise, $\boldsymbol{\alpha} = (\alpha_1,\alpha_2,\alpha_3)$ and β are the Hermitian 4×4 Dirac matrices defined by the commutation relations ($i,j = 1,2,3$),

0022-2488/88/010249-09$02.50

$$\alpha_i\alpha_j + \alpha_j\alpha_i = 2\delta_{ij}\mathbb{1}, \quad \alpha_i\beta + \beta\alpha_i = 0, \quad \beta^2 = 1. \tag{2.3}$$

We obtain the form (1.1) for H_0 by choosing Dirac's representation for $\boldsymbol{\alpha}$ and β:

$$H_0 = \begin{pmatrix} mc^2 & c\boldsymbol{\sigma}\cdot\mathbf{P} \\ c\boldsymbol{\sigma}\cdot\mathbf{P} & -mc^2 \end{pmatrix} \equiv h(\mathbf{P}). \tag{2.4}$$

Here $\boldsymbol{\sigma} = (\sigma_1,\sigma_2,\sigma_3)$ are the Pauli matrices, $D = D^* = c\boldsymbol{\sigma}\cdot\mathbf{P}$ is defined in $\mathscr{H}_+ = \mathscr{H}_- = L^2(\mathbb{R}^3)^2$.

In this case, U and V are easily determined directly. By a Fourier transformation $\mathscr{F}$, the matrix differential operator (4.4) becomes multiplication by the Hermitian matrix-valued function $h(\mathbf{p})$. For each $\mathbf{p}\in\mathbb{R}^3$ we can diagonalize this by a unitary matrix $u(\mathbf{p})$. Defining

$$\mathscr{W} = u(\cdot)\circ\mathscr{F} \tag{2.5}$$

we obtain for $\Psi\in\mathscr{D}(H_0)$

$$(\mathscr{W}H_0\Psi)(\mathbf{p}) = \begin{pmatrix} (c^2p^2 + m^2c^4)^{1/2}\mathbb{1} & 0 \\ 0 & -(c^2p^2 + m^2c^4)^{1/2}\mathbb{1} \end{pmatrix} \times (\mathscr{W}\Psi)(\mathbf{p}), \tag{2.6}$$

which shows that the unitary transformation diagonalizing (2.4) is given by

$$U_0 := \mathscr{F}^{-1}\mathscr{W}. \tag{2.7}$$

For other methods of deriving the Foldy–Wouthuysen transformation (2.7) see, e.g., Ref. 1.

B. The Dirac operator in external fields

There are various other situations in which the Dirac operator has the form (1.1). For a charged particle in a magnetic field $\mathbf{B}(\mathbf{x}) = \operatorname{rot}\mathbf{A}(\mathbf{x})$, $\mathbf{A}\in C^1(\mathbb{R}^3)$ (setting $c = 1$), we have

$$D = \boldsymbol{\sigma}\cdot(\mathbf{P} - \mathbf{A}(\mathbf{x})), \quad M_+ = M_- = m. \tag{2.8}$$

[Note that for $\mathbf{A}$ as above, D is symmetric and thus closable on $C_0^\infty(\mathbb{R}^3)^2$.]

If the particle has an anomalous electric moment δ, then

$$D = \boldsymbol{\sigma}\cdot(\mathbf{P} - \mathbf{A} - i\delta\mathbf{B}). \tag{2.9}$$

Similarly, for a neutron with anomalous magnetic moment μ in an electric field $\mathbf{E}(\mathbf{x})$ (cf. Ref. 7),

$$D = \boldsymbol{\sigma}\cdot(\mathbf{P} - i\mu\mathbf{E}). \tag{2.10}$$

If $\mathbf{E}(\mathbf{x}) = \mathbf{E}(r)$, $r = |\mathbf{x}|$, then also the radial Dirac operator takes the form (1.1), with

$$D = \frac{d}{dr} - \frac{\kappa}{r} + \mu E(r) \tag{2.11}$$

defined in $\mathscr{H}_+ = \mathscr{H}_- = L^2((0,\infty),dr)$. Even for extremely singular $E(r)$ the operator (2.11) is closable on $C_0^\infty((0,\infty))$ (cf. Ref. 18 and the references therein).

For the Dirac operator in an external scalar field $V(\mathbf{x})$ we use the following "supersymmetric" representation of the Dirac matrices

$$\boldsymbol{\alpha} = \begin{pmatrix} 0 & \boldsymbol{\sigma} \\ \boldsymbol{\sigma} & 0 \end{pmatrix}, \quad \beta = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix} \tag{2.12}$$

to obtain the Dirac operator in the form (1.1) with

$$D = \boldsymbol{\sigma}\cdot\mathbf{p} + i(m + V(\mathbf{x})), \quad M_+ = M_- = 0. \tag{2.13}$$

Note that in the examples (2.9)–(2.13) D is not self-adjoint, and not even symmetric.

C. The Klein–Gordon equation

For a spin-0 particle with mass $m > 0$ in an external magnetic field the Klein–Gordon equation reads

$$\left\{\frac{\partial^2}{\partial t^2} + (\mathbf{P} - \mathbf{A}(\mathbf{x}))^2 + m^2\right\}u(\mathbf{x},t) = 0. \tag{2.14}$$

Writing this as a first-order system we obtain

$$i\frac{d}{dt}\Psi(t) = H\Psi(t), \quad \Psi(t) = \begin{pmatrix} u(\cdot,t) \\ i\,\partial u(\cdot,t)/\partial t \end{pmatrix}, \tag{2.15}$$

$$H = \begin{pmatrix} 0 & \mathbb{1} \\ T & 0 \end{pmatrix}, \quad T = (p - A)^2 + m^2. \tag{2.16}$$

We shall define Hilbert spaces $\mathscr{H}_+$, $\mathscr{H}_-$ such that (2.16) obtains the form (1.1) in a natural way.

Under suitable conditions on $\mathbf{A}(\mathbf{x})$ (Ref. 19) T is essentially self-adjoint on $C_0(\mathbb{R}^3)$. Its closure, also denoted by T is strictly positive on $\mathscr{D}(T)\subset L^2(\mathbb{R}^3)$. The vector space

$$\mathscr{H}_+ := \mathscr{D}(T^{1/2}) \tag{2.17}$$

is a Hilbert space with the scalar product

$$(u,v)_{\mathscr{H}_+} := (T^{1/2}u, T^{1/2}v) \tag{2.18}$$

(with the ordinary L^2-scalar product on the right-hand side). Furthermore, define

$$\mathscr{H}_- := L^2(\mathbb{R}^3). \tag{2.19}$$

Note that $\Psi(t)$ has finite norm in $\mathscr{H}$ iff u has finite field energy,

$$\|\Psi\|^2 = \int d^3x\left\{|(-i\nabla - \mathbf{A})u|^2 + m^2|u|^2 + \left|\frac{\partial}{\partial t}u\right|^2\right\}. \tag{2.20}$$

In (2.16) the operator

$$\mathbb{1}: \mathscr{D}(T^{1/2})\subset\mathscr{H}_- \to \mathscr{H}_+ \tag{2.21}$$

is the adjoint of

$$T: \mathscr{D}(T)\subset\mathscr{H}_+ \to \mathscr{H}_-. \tag{2.22}$$

Thus we can set $D := T$, $D^* = \mathbb{1}$ [as in (2.21)], $M_+ = 0$ (on $\mathscr{H}_+$), and $M_- = 0$ (on $\mathscr{H}_-$).

D. Temple's operator

In polar coordinates, the Dirac operator in a Coulomb field γ/r reads

$$H = \boldsymbol{\alpha}\cdot\mathbf{p} + \beta m + \frac{\gamma}{r} = -i\frac{\boldsymbol{\alpha}\cdot\mathbf{x}}{r}\left\{\frac{\partial}{\partial r} + \frac{1}{r} - \frac{\Gamma}{r}\right\} + \beta m, \tag{2.23}$$

where we have introduced Temple's operator (cf. Ref. 20 and the references therein)

$$\Gamma = (\boldsymbol{\sigma}\cdot\mathbf{L} + 1) - \frac{i\gamma}{r}\boldsymbol{\alpha}\cdot\mathbf{x} = \begin{pmatrix} \boldsymbol{\sigma}\cdot\mathbf{L} + 1 & -(i\gamma/r)\boldsymbol{\sigma}\cdot\mathbf{x} \\ -(i\gamma/r)\boldsymbol{\sigma}\cdot\mathbf{x} & \boldsymbol{\sigma}\cdot\mathbf{L} + 1 \end{pmatrix} \tag{2.24}$$

($\mathbf{L} = \mathbf{x}\wedge\mathbf{p}$). This operator is not self-adjoint. Nevertheless, since the two summands in (2.24) anticommute and since

Γ^2 is self-adjoint and positive, our theory generalizes to this situation (cf. the remarks at the end of Sec. V). Here Γ is diagonalized in order to bring the Dirac Coulomb equation to a two-component form which admits an algebraic solution of the relativistic Coulomb problem.[21,22]

III. MAIN RESULTS

We start by giving the precise definitions in the language of supersymmetry. The connection with the problem of bringing (1.1) to the diagonal form will become clear when we introduce the "standard representation."

Let $\mathscr{H}$ be a Hilbert space with a self-adjoint involution, i.e., an operator τ satisfying

$$\tau^*\tau = \tau\tau^* = \tau^2 = 1 \,. \tag{3.1}$$

A self-adjoint operator Q on $\mathscr{D}(Q)$ satisfying

$$\tau Q + Q\tau = 0 \tag{3.2}$$

(in the sense of quadratic forms on $\mathscr{D}(Q)\times\mathscr{D}(Q)$, cf. Ref. 12) is called a supercharge with respect to τ. The operator Q^2 is usually called a Hamiltonian with supersymmetry. By the spectral theorem it is densely defined, self-adjoint, and positive. Let M be a symmetric operator which is relatively bounded with respect to Q and has Q bound less than one. We further assume that M commutes with τ and anticommutes with Q, i.e., (again in the quadratic form sense),

$$\tau M - M\tau = 0 \,, \tag{3.3}$$

$$QM + MQ = 0 \,. \tag{3.4}$$

(For an antisymmetric M which occurs, e.g., in Sec. II D, cf. the remarks at the end of Sec. V.)

For any pair of operators Q and M defined as above there are associated operators Q' and M' given by

$$Q' := iQ\tau, \quad M' := M\tau \,. \tag{3.5}$$

They satisfy the relations (3.2), (3.3), and

$$Q'^2 = Q^2, \quad M'^2 = M^2 \,, \tag{3.6}$$

$$Q'Q + QQ' = 0, \quad M'M - MM' = 0 \,, \tag{3.7}$$

and, instead of (3.4),

$$Q'M' - M'Q' = 0 \,. \tag{3.8}$$

Our aim is to study the abstract Dirac operator

$$H := Q + M \,, \tag{3.9}$$

which by the Kato–Rellich theorem is self-adjoint on $\mathscr{D}(H) = \mathscr{D}(Q)$. From (3.4) we conclude

$$H^2 = Q^2 + M^2 \text{ on } \mathscr{D}(H^2) = \mathscr{D}(Q^2) \,. \tag{3.10}$$

Note that H^2 commutes both with M and Q.

In order to formulate our results we need some additional notation. Define, for $\lambda \neq 0$,

$$a_\pm(\lambda^2, m) := 2^{-1/2}\{1 \pm m(\lambda^2 + m^2)^{-1/2}\}^{1/2} \,. \tag{3.11}$$

Functions of self-adjoint operators are defined as usual via the spectral theorem. By an abuse of notation we shall also write

$$a_\pm(Q^2, M) := 2^{-1/2}\{1 \pm M(Q^2 + M^2)^{-1/2}\}^{1/2} \quad \text{on Ker } Q^\perp \,. \tag{3.12}$$

Note that $M(Q^2 + M^2)^{-1/2}$ extends to a bounded operator on Ker $Q^\perp$. We denote

$$|T| := (T^2)^{1/2} \text{ on } \mathscr{D}(T) = \mathscr{D}(|T|) \,, \tag{3.13}$$

$$\operatorname{sgn} T := \begin{cases} T|T|^{-1} = |T|^{-1}T & \text{on Ker } T^\perp \,, \quad (3.14a) \\ 0 \quad \text{on Ker } T \,, & \quad (3.14b) \end{cases}$$

for any self-adjoint T. In (3.14a) the expressions on the right-hand side are meant as closures of the densely defined bounded operators.

Our first result states that the abstract Dirac operator H is unitarily equivalent to an operator $H_{\rm FW}$ that commutes with τ, anticommutes with Q, and satisfies $H_{\rm FW}^2 = H^2$.

Theorem 1: There is a unitary operator U in $\mathscr{H}$ such that (on Ker $Q^\perp$)

$$UHU^* = |H|\tau =: H_{\rm FW} \,. \tag{3.15}$$

Here U depends on Q', M' defined in (1.6) as follows:

$$U = a_+(Q'^2, M') + ia_-(Q'^2, M')\operatorname{sgn} Q' \,. \tag{3.16}$$

The abstract Dirac operator is not a supercharge itself, but it is unitarily equivalent to one. This is the content of our second result.

Theorem 2: There is a unitary operator V in Ker $Q^\perp \subset \mathscr{H}$ such that

$$VHV^* = |H|\operatorname{sgn} Q =: H_{\rm CT} \,. \tag{3.17}$$

It is given by

$$V = 2^{-1/2}(1 - i\operatorname{sgn} Q')U \,. \tag{3.18}$$

Obviously, $H_{\rm CT}$ commutes with Q and anticommutes with τ. Furthermore we have $H_{\rm CT}^2 = H^2$ and thus H^2 is the supersymmetric Hamiltonian associated to the supercharge $H_{\rm CT}$. The proof of Theorems 1 and 2 together with further results will be given in Sec. III.

Theorem 1 solves the above mentioned problem of diagonalizing the operator-valued matrix (1.1). In order to see the connection we define the following standard representation, which is completely equivalent to the general framework.

Let $\mathscr{H}_+$ (resp. $\mathscr{H}_-$) be Hilbert spaces and D: $\mathscr{D}_+ \subset \mathscr{H}_+ \to \mathscr{H}_-$ be a densely defined closed linear operator. Then the adjoint operator D^* from $\mathscr{H}_-$ to $\mathscr{H}_+$ exists and is defined on a dense set $\mathscr{D}_- \subset \mathscr{H}_-$. In the Hilbert space

$$\mathscr{H} := \mathscr{H}_+ \oplus \mathscr{H}_- \tag{3.19}$$

(the direct sum) the operator

$$Q := \begin{pmatrix} 0 & D^* \\ D & 0 \end{pmatrix} \text{ on } \mathscr{D}(Q) = \mathscr{D}_+ \oplus \mathscr{D}_- \tag{3.20}$$

is self-adjoint. The equivalence to the supersymmetric framework can be seen by identifying $\mathscr{H}_\pm$ with the eigenspaces of τ belonging to the eigenvalues ± 1, respectively. Then any supercharge with respect to τ is represented by a matrix operator of the form (3.20). The operator M is now given by

$$M = \begin{pmatrix} M_+ & 0 \\ 0 & -M_- \end{pmatrix}, \tag{3.21}$$

where the M_+ (resp. M_-) are symmetric operators in $\mathscr{H}_+$

(resp. $\mathscr{H}_-$), bounded relative to D (resp. D^*) (with relative bound less than 1), and such that

$$D^*M_- = M_+D^* \quad \text{on } \mathscr{D}_+ \times \mathscr{D}_- , \tag{3.22a}$$

$$DM_+ = M_-D \quad \text{on } \mathscr{D}_- \times \mathscr{D}_+ , \tag{3.22b}$$

in the quadratic form sense. Equation (3.22) is equivalent to (3.4) and the relation (3.3) is automatically satisfied by (3.21).

The abstract Dirac operator reads in the standard representation [cf. (1.1)]

$$H = \begin{pmatrix} M_+ & D^* \\ D & -M_- \end{pmatrix} \quad \text{on } \mathscr{D}(H) = \mathscr{D}_+ \oplus \mathscr{D}_- . \tag{3.23}$$

Next define the self-adjoint and positive operators

$$Q_+ := (D^*D)^{1/2} \quad [\text{resp. } Q_- := (DD^*)^{1/2}] \tag{3.24}$$

on $\mathscr{D}_+$ (resp. $\mathscr{D}_-$). Then

$$\begin{aligned} H^2 = \begin{pmatrix} H^2_+ & 0 \\ 0 & H^2_- \end{pmatrix} \quad &\text{on } \mathscr{D}(H^2) \\ &= \mathscr{D}(Q^2_+) \oplus \mathscr{D}(Q^2_-) , \end{aligned} \tag{3.25}$$

where

$$H_\pm = (Q^2_\pm + M^2_\pm)^{1/2} \quad \text{on } \mathscr{D}_\pm . \tag{3.26}$$

Finally, the normal forms of the abstract Dirac operator are given by

$$\begin{aligned} H_{\rm FW} &= \begin{pmatrix} H_+ & 0 \\ 0 & -H_- \end{pmatrix}, \\ H_{\rm CT} &= \begin{pmatrix} 0 & D^*Q^{-1}_- H_- \\ DQ^{-1}_+ H_+ & \end{pmatrix}. \end{aligned} \tag{3.27}$$

Further details on the unitary transformations U and V are given in Sec. V. In the Appendix, for the sake of concreteness, we provide a list of the explicit expressions for $Q^2_\pm$ obtained for the examples in Sec. II. Further details on the transformation U at V are given in Sec. V.

If one wishes to do calculations within the standard representation one needs a commutation formula (4.15) for closed operators which has been proved (in a slightly different form) by Deift.[11] In Sec. IV we include a simple new proof of this formula which is quite natural from the supersymmetric point of view.

In the study of the nonrelativistic limit of the Dirac equation the abstract form (3.23) of the Dirac operator has been useful in revealing the underlying structure of the problem.[10,16] Here Q^2_+ (resp. Q^2_-) appear as the nonrelativistic limits of the Dirac operator

$$H(c,m) := \begin{pmatrix} mc^2 & cD^* \\ cD & -mc^2 \end{pmatrix} \tag{3.28}$$

with rest energy $mc^2 \cdot \mathbb{1}$ subtracted (resp. added). One has, for example, the result[16]

$$\begin{aligned} &\operatorname*{n-lim}_{c\to\infty} \{H(c,m) - mc^2 - z\}^{-1} \\ &\quad = \begin{pmatrix} \{(2m)^{-1}Q^2_+ - z\}^{-1} & 0 \\ 0 & 0 \end{pmatrix}. \end{aligned} \tag{3.29}$$

In the applications c is the velocity of light and m the mass of the particle. From (3.29) it is easy to see that in the norm resolvent sense

$$\lim_{c\to\infty} (H(c,m) - H_{\rm FW}(c,m)) = 0 . \tag{3.30}$$

On the other hand, in the "extreme relativistic" limit $m \to 0$, we have

$$\lim_{m\to 0} (H(c,m) - H_{\rm CT}(c,m)) = 0 . \tag{3.31}$$

IV. NELSON'S TRICK

The statement that the operators D^*D and DD^* are densely defined if D is densely defined and closed is a well known theorem of von Neumann. The proof usually found in the textbooks is rather complicated (cf. Ref. 23, § V.3.7). In the language of Sec. I this theorem follows almost trivially from the spectral theorem for self-adjoint operators and the fact that

$$Q := \begin{pmatrix} 0 & D^* \\ D & 0 \end{pmatrix} \tag{4.1}$$

is self-adjoint if and only if D is closed. Just note that by the spectral theorem Q^2 is densely defined on

$$\begin{aligned} \mathscr{D}(Q^2) &= \{ f \in \mathscr{D}(Q) \,|\, Qf \in \mathscr{D}(Q) \} \\ &= \mathscr{D}(D^*D) \oplus \mathscr{D}(DD^*) . \end{aligned} \tag{4.2}$$

This argument is due to Nelson (unpublished). We shall use it as a method of proving results for closed operators D by proving analogous results for the self-adjoint operators Q (where the spectral theorem does the hard work).

For example, the statement

$$Q = |Q| \operatorname{sgn} Q = \operatorname{sgn} Q \, |Q| , \tag{4.3}$$

which is immediate from the spectral theorem and the definitions (3.13) and (3.14), is equivalent to the polar decomposition theorem for closed operators (cf. Ref. 23, § VI.2.7): If D is densely defined and closed, then

$$D = Q_-S = SQ_+ , \tag{4.4}$$

where the $Q_\pm$ are defined as in (3.24) and

$$S := \begin{cases} Q^{-1}_- D = DQ^{-1}_+ & \text{on } \operatorname{Ker} D^\perp , \quad (4.5\text{a}) \\ 0 \quad \text{on } \operatorname{Ker} D . & (4.5\text{b}) \end{cases}$$

We have

$$\operatorname{sgn} Q = \begin{pmatrix} 0 & S^* \\ S & 0 \end{pmatrix}. \tag{4.6}$$

The formula

$$\operatorname{Ker} Q = \operatorname{Ker} Q^2 = \operatorname{Ran} Q^\perp \tag{4.7}$$

is equivalent to

$$\operatorname{Ker} D = \operatorname{Ker} D^*D = \operatorname{Ran} D^{*\perp} , \tag{4.8a}$$

$$\operatorname{Ker} D^* = \operatorname{Ker} DD^* = \operatorname{Ran} D^\perp . \tag{4.8b}$$

From (4.4) and (4.5) we obtain immediately

$$Q^2_- S = SQ^2_+ \tag{4.9}$$

and therefore for the spectra

$$\sigma(Q_+)\setminus\{0\} = \sigma(Q_-)\setminus\{0\}. \tag{4.10}$$

Similarly, with the help of (3.22) we obtain for the spectra of the self-adjoint operators $H_\pm$ defined in (3.26)

$$\sigma(H_+)\setminus\{0\} = \sigma(H_-)\setminus\{0\}. \tag{4.11}$$

By Nelson's trick the formula

$$\mathbb{1} + z(Q^2 - z)^{-1} = Q^2(Q^2 - z)^{-1} = Q(Q^2 - z)^{-1}Q \tag{4.12}$$

may be rewritten as

$$\mathbb{1} + z(DD^* - z)^{-1} = D(D^*D - z)^{-1}D^*, \tag{4.13a}$$

$$\mathbb{1} + z(D^*D - z)^{-1} = D^*(DD^* - z)^{-1}D. \tag{4.13b}$$

For bounded, measurable f we have

$$Qf(Q^2) = f(Q^2)Q \quad \text{on } (Q), \tag{4.14}$$

which becomes

$$D^*f(DD^*) = f(D^*D)D^*, \tag{4.15a}$$

$$Df(D^*D) = f(DD^*)D, \tag{4.15b}$$

on $\mathscr{D}_+$ (resp. $\mathscr{D}_-$). Equations (4.13) and (4.15) (for $f=$ resolvant) have been proved first by Deift[11] using a different method. Since also the bounded operator $M'(Q^2 + M^2)^{-1/2}$ commutes with Q we easily derive in the same way the following relations:

$$D^*a_\pm(Q^2_-, M_-) = a_\pm(Q^2_+, M_+)D^*, \tag{4.16a}$$

$$Da_\pm(Q^2_+, M_+) = a_\pm(Q^2_-, M_-)D, \tag{4.16b}$$

which again hold on $\mathscr{D}_+$ (resp. $\mathscr{D}_-$).

V. THE TRANSFORMATIONS *U* AND *V*

In this section we shall prove Theorems 1 and 2 together with some further results on U and V (Theorem 3). The following remarks on eigenvectors are meant also as a motivation for the ansatz (5.11).

Given eigenvectors of the "nonrelativistic" operators $Q^2_\pm$, $M_\pm$ it is easy to determine eigenvectors of the abstract Dirac operator. Let $u \in \mathscr{H}_+$ be a simultaneous eigenvector of Q^2_+ and M_+ such that for some $\lambda > 0$, $m \in \mathbf{R}$,

$$\|u\| = 1, \quad Q^2_+ u = \lambda^2 u, \quad M_+ u = mu. \tag{5.1}$$

Similarly, let $v \in \mathscr{H}_-$ be an eigenvector of Q^2_- and M_- with the same eigenvalues,

$$\|v\| = 1, \quad Q^2_- v = \lambda^2 v, \quad M_- v = mv. \tag{5.2}$$

For a given u we can always find such a vector v as long as $\lambda \neq 0$ (i.e., $u \notin \text{Ker } D$), take

$$v := \|Du\|^{-1}Du. \tag{5.3}$$

In view of (1.25) it seems natural to make the ansatz

$$H\begin{pmatrix} u \\ k \quad Du \end{pmatrix} = E\begin{pmatrix} u \\ k \quad Du \end{pmatrix}$$

from which E and k can indeed be determined as functions of λ^2 and m. More precisely, we define

$$f := n\begin{pmatrix} u \\ k \quad Du \end{pmatrix}, \tag{5.4}$$

$$g := n\begin{pmatrix} -kD^*v \\ v \end{pmatrix}, \tag{5.5}$$

with

$$k \equiv k(\lambda^2, m) = (1/\lambda^2)\{(\lambda^2 + m^2)^{1/2} - m\}. \tag{5.6}$$

Here n is a normalization factor such that

$$\|f\| = \|g\| = 1. \tag{5.7}$$

A simple calculation shows that

$$n \equiv n(\lambda^2, m) = \{1 + \lambda^2 k^2\}^{-1/2} = a_+(\lambda^2, m), \tag{5.8}$$

and

$$n(\lambda^2, m)k(\lambda^2, m) = a_-(\lambda^2, m)/|\lambda| \tag{5.9}$$

[with the functions $a_\pm$ defined in (3.11)].

Here f and g are orthogonal, $(f,g) = 0$, and satisfy

$$Hf = (\lambda^2 + m^2)^{1/2}f, \tag{5.10a}$$

$$Hg = -(\lambda^2 + m^2)^{1/2}g. \tag{5.10b}$$

Now we are ready to define in the spirit of linear algebra the operator W^* (which is, in a sense, the "matrix of eigenvectors") such that $W^*\binom{u}{0} = f$ and $W^*\binom{0}{v} = g$:

$$W^* \equiv \begin{pmatrix} a + (Q^2_+, M_+) & -D^*Q_-^{-1}a_-(Q^2_-, M_-) \\ DQ_+^{-1}a_-(Q^2_+, M_+) & a_+(Q^2_-, M_-) \end{pmatrix} \tag{5.11a}$$

$$= a_+(Q^2, M') - iQ'|Q|^{-1}a_-(Q^2, M') \tag{5.11b}$$

$$= a_+(Q'^2, M') - i(\text{sgn } Q')a_-(Q'^2, M'). \tag{5.11c}$$

Unitarity of U^* is easily seen from $a^2_+ + a^2_- = 1$. Thus we set $U := W^{*-1} = a_+ + i(\text{sgn } Q')a_-$ to obtain the operator defined in Theorem 1. Once having the explicit form of U it is easy to verify (3.15),

$$UHU^* = (a_+ + \tau a_- \text{ sgn } Q)H(a_+ - \tau a_- \text{ sgn } Q)$$

$$= (a^2_+ + 2\tau a_+ a_- \text{ sgn } Q - a^2_-)H,$$

where we have used

$$H\tau \text{ sgn } Q = -\tau(\text{sgn } Q)H \tag{5.12}$$

and the fact that $a_\pm(Q'^2, M')$ commutes with both Q and M. Using

$$a^2_+ - a^2_- = \tau M|H|^{-1}, \quad 2a_+a_- = |Q|\,|H|^{-1} \tag{5.13}$$

and (4.3) we finally obtain the result of Theorem 1. □

For $M = 0$ we have $a_\pm(Q'^2, 0) = 2^{-1/2}$ on Ker $Q^\perp$ and therefore U specializes to

$$\widetilde{U} := 2^{-1/2}(\mathbb{1} + i \text{ sgn } Q') = \frac{1}{\sqrt{2}}\begin{pmatrix} \mathbb{1} & S^* \\ -S & \mathbb{1} \end{pmatrix}, \tag{5.14}$$

where S is the operator defined in (4.5). Now it is easy to see that

$$\widetilde{U}^* H_{\text{FW}} \widetilde{U} = \begin{pmatrix} 0 & S^*|H| \\ S|H| & 0 \end{pmatrix}, \tag{5.15}$$

which is just the operator H_{CT} defined in Theorem 2. Therefore the unitary operator

$$V := \widetilde{U}^* U$$

satisfies $VHV^* = H_{\text{CT}}$, which proves Theorem 2. □

For any self-adjoint operator A define the Cayley transform C as the unitary operator

$$C := (i - A)(i + A)^{-1} = \exp\{2i \arctan A\}. \tag{5.16}$$

Conversely, if C is the Cayley transform of a self-adjoint operator A, then A is given by

$$\mathscr{D}(A) = \operatorname{Ran}(1 + C), \quad A = i(1 - C)(1 + C)^{-1}. \tag{5.17}$$

Theorem 3: Let M be bounded and M^2 be strictly positive. Then the unitary operator U^4 is the Cayley transform of QM^{-1} and on Ker $Q^{\perp}$ the operator V^4 is the Cayley transform of $-MQ^{-1}$. In this case one has on Ker $Q^{\perp}$ the representations

$$U = \exp\{(i/2)\arctan QM^{-1}\}, \tag{5.18}$$

$$V = \exp\{-(i/2)\arctan MQ^{-1}\} \quad \text{on } (\operatorname{Ker} Q)^{\perp}. \tag{5.19}$$

Proof: Here $M = \tau M'$ is bounded and symmetric and hence self-adjoint. From $M^2 > 0$ we see that M has a gap around 0 in this spectrum and thus M is bijective. Here M^{-1} is also bounded and commutes with Q. Therefore QM^{-1} is defined on $\mathscr{D}(Q)$ and self-adjoint. Moreover it is injective in the Hilbert space $(\operatorname{Ker} Q)^{\perp}$ and has dense range in $(\operatorname{Ker} Q)^{\perp}$. We can therefore define the inverse MQ^{-1} as a self-adjoint operator in $(\operatorname{Ker} Q)^{\perp}$. A little calculation shows that for $\lambda \in \mathbb{R}$ and arbitrary m

$$\{a_+(\lambda^2,m) + i(\lambda/|\lambda|)a_-(\lambda^2,m)\}^4 = (i - \lambda/m)/(i + \lambda/m). \tag{5.20}$$

The result (3.18) now follows from (3.16) with $A = Q'M'^{-1} = QM^{-1}$. Since $\widetilde{U}^4 = -1$ (and since $\widetilde{U}$ commutes with U) we have

$$U^4 = -V^4.$$

Equation (3.19) follows from the observation that for $\lambda \neq 0$

$$(i - \lambda/m)/(i + \lambda/m) = -(i + m/\lambda)/(i - m/\lambda). \quad \square$$

If D is self-adjoint and $M_+ = M_-$ then we do not need condition (3.22) (e.g., in case of the Dirac operator with a Lorentz-scalar potential). In fact, with

$$T = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & i \\ i & 1 \end{pmatrix}$$

and $\widetilde{D} := D + iM_+$ we have

$$T\begin{pmatrix} M_+ & D \\ D & -M_+ \end{pmatrix}T^{-1} = \begin{pmatrix} 0 & \widetilde{D}^* \\ \widetilde{D} & 0 \end{pmatrix}.$$

Therefore we can apply our theory (in particular the transformation $\widetilde{U}$ [cf. (5.14)]) to obtain the diagonal form. The unitary operator T just transforms the standard representation into one where τ has the form

$$\tau = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}.$$

Finally we consider Temple's operator (cf. Sec. II D). It is not symmetric, but nevertheless the preceding theory is easily adapted to this situation. We simply identify $\mathscr{H}_{\pm} = L^2(\mathbb{R}^3)^2$, $\tau = \beta$, $M = \boldsymbol{\sigma}\cdot\mathbf{L} + 1 \equiv \beta K$, $Q = (-i\gamma/r)\boldsymbol{\alpha}\cdot\mathbf{x}$ and note that

$$H^2 = Q^2 + M^2 = K^2 - \gamma^2/c^2 = \Gamma^2$$

is self-adjoint and strictly positive as long as $|\gamma|/c < 1$ (which is precisely the condition for the essential self-adjointness of the Coulomb problem). According to (3.12) we can define

$$a_{\pm} = (1/\sqrt{2})(1 \pm K/\sqrt{\Gamma^2})^{1/2}$$

as a function of the bounded self-adjoint operator $K/\sqrt{\Gamma^2}$. The transformation U, as defined in Theorem 1, reads

$$U = a_+ - \beta a_- \boldsymbol{\alpha}\cdot(\mathbf{x}/r)\operatorname{sgn}\gamma.$$

This operator is not unitary, but since all algebraic relations for the a_+ and a_- remain unaffected we obtain

$$U^{-1} = a_+ + \beta a_- \boldsymbol{\alpha}\cdot(\mathbf{x}/r)\operatorname{sgn}\gamma$$

and

$$U\Gamma U^{-1} = \beta\sqrt{\Gamma^2}.$$

Of course V is not unitary either. As an analog of Theorem 3 we obtain in this case the representations

$$U = \exp(\tfrac{1}{2}\operatorname{artanh}(iQM^{-1})),$$

$$V = \exp(-\tfrac{1}{2}\operatorname{artanh}(-iMQ^{-1})).$$

VI. SCATTERING THEORY

In this section we assume for simplicity

$$M = m\tau, \quad m > 0. \tag{6.1}$$

Let Q and Q_0 be supercharges with respect to τ and consider the pair of abstract Dirac operators

$$H = Q + m\tau, \quad H_0 = Q_0 + m\tau. \tag{6.2}$$

We are interested in the existence of the wave operator[24]

$$\Omega_{\pm}(H,H_0) := \operatorname*{s-lim}_{t\to\pm\infty} e^{iHt}e^{-iH_0t}P_{\text{a.c.}}(H_0), \tag{6.3}$$

where $P_{\text{a.c.}}(\cdot)$ denotes the spectral projection to the absolutely continuous subspace of the indicated operator.

Often it is easier to show existence of the nonrelativistic wave operators $\Omega_{\pm}(Q^2,Q_0^2)$. Consider, for example, the Dirac operator in a magnetic field $\mathbf{B}$ (cf. Sec. II) satisfying $(\mathbf{x}\in\mathbb{R}^3)$

$$|\mathbf{B}(\mathbf{x})| \leqslant \text{const}(1 + |\mathbf{x}|)^{-3/2-\delta}, \quad \delta > 0 \tag{6.4}$$

(B has finite field energy, $\int B^2 < \infty$). Choosing the transversal gauge

$$\mathbf{A}(\mathbf{x}) = \mathbf{G}(\mathbf{x}) \wedge \mathbf{x}, \tag{6.5}$$

$$G(\mathbf{x}) = \int_0^1 ds\, s\mathbf{B}(\mathbf{x}s), \tag{6.6}$$

the vector potential $\mathbf{A}$ decays like

$$|\mathbf{A}(\mathbf{x})| \leqslant \text{const}(1 + |\mathbf{x}|)^{-1/2-\delta}. \tag{6.7}$$

In general, the decay properties of the vector potential cannot be improved by choosing a different gauge. Therefore, Schrödinger and Dirac operators contain long-range terms and the existence of the wave operators cannot be shown by standard techniques. (Even the classical scattering theory is "long range," i.e., the motion of the particles is not approximated by the free motion for large times.) In the nonrelativistic case, however, we can proceed as follows.[25] We have

$$\begin{aligned} Q^2 &= \mathbf{P}^2 - 2\mathbf{A}\cdot\mathbf{P} + i\operatorname{div}\mathbf{A} + \mathbf{A}^2 - \boldsymbol{\sigma}\cdot\mathbf{B}, \\ Q_0^2 &= \mathbf{P}^2 = -\Delta. \end{aligned} \tag{6.8}$$

In the transversal gauge the long-range term $\mathbf{A}\cdot\mathbf{P}$ can be written as

$$\mathbf{A}(\mathbf{x})\cdot\mathbf{P}=\mathbf{G}(\mathbf{x})\cdot\mathbf{L}\,, \tag{6.9}$$

$\mathbf{L}$ being the orbital angular momentum. A simple nonstationary phase argument (cf. Ref. 24, Appendix 1 to Sec. XI.3) shows for Ψ in a suitable dense domain

$$\|(Q^2-Q_0^2)\exp(-iQ_0^2t)\Psi\|\leqslant\text{const}(1+|t|)^{-1-2\delta}\,. \tag{6.10}$$

We have used that $\mathbf{L}$ commutes with Q_0^2 and that div $\mathbf{A}$, A^2, $\boldsymbol{\sigma}\cdot\mathbf{B}$, and G are of short range. Applying the Cook argument (cf. Ref. 24, Sec. XI.4), existence of $\Omega_\pm(Q^2,Q_0^2)$ follows immediately. In the relativistic case the long-range term is

$$H-H_0=\boldsymbol{\alpha}\cdot\mathbf{A}(\mathbf{x})\,. \tag{6.11}$$

Unlike (6.10), the expression

$$\|(H-H_0)\exp\{-iH_0t\}\Psi\| \tag{6.12}$$

does not decay fast enough. Due to the "*Zitterbewegung*" $\boldsymbol{\alpha}$ gives no contribution to the decay and the Cook argument has to be modified in an existence proof.[26,27] (See also Ref. 28 for a discussion of *Zitterbewegung* in relativistic scattering theory.)

Here we shall give an abstract criterion for existence of (6.3) where only little more has to be checked than what is needed for the existence of the nonrelativistic wave operators $\Omega_\pm(Q^2,Q_0^2)$. In the following $F(\cdot)$ denotes the spectral projection of the operator to the part of the spectrum as indicated in the parentheses; $\mathscr{H}_{\text{a.c.}}(\cdot)$ is the absolutely continuous subspace.

Theorem 4: Let H,H_0 be given as in (6.2), $H_{\text{FW}}^0=|H_0|\tau$. Assume that for all $0<a<b<\infty$ and for Ψ in some dense subset of $F(a<Q_0^2<b)\,\mathscr{H}_{\text{a.c.}}(Q_0^2)$,

$$\|(Q^2-Q_0^2)\exp(-iQ_0^2t)\Psi\|\leqslant\text{const}(1+|t|)^{1+\delta} \tag{6.13}$$

and

$$\|(Q-Q_0)\exp(-iH_{\text{FW}}^0t)\Psi\|\to0,\quad\text{as }|t|\to\infty\,. \tag{6.14}$$

Then $\Omega_\pm(H,H_0)$ exist, and

$$\begin{aligned}\Omega_\pm(H,H_0)=\Omega_\pm(Q^2,Q_0^2)F(H_0>0)&\\+\Omega_\mp(Q^2,Q_0^2)F(H_0<0)\,.&\end{aligned} \tag{6.15}$$

Proof: Equation (6.13) implies existence of $\Omega_\pm(Q^2,Q_0^2)$ by the Cook argument. By the invariance principle (Ref. 24, Appendix 3 to Sec. XI.3) we conclude existence of $\Omega_\pm(H_{\text{FW}},H_{\text{FW}}^0)$, and

$$\begin{aligned}\Omega_\pm(H_{\text{FW}},H_{\text{FW}}^0)=\tfrac12\Omega_\pm(Q^2,Q_0^2)(1+\tau)&\\+\tfrac12\Omega_\mp(Q^2,Q_0^2)(1-\tau)\,.&\end{aligned} \tag{6.16}$$

We have to show that for all $\Psi\in\mathscr{H}_{\text{a.c.}}(H_0)$ we can find $\Phi_\pm\in\mathscr{H}$ such that

$$0=\lim_{t\to\pm\infty}\|e^{iHt}e^{-iH_0t}\Psi-\Phi_\pm\| \tag{6.17a}$$

$$\leqslant\lim_{t\to\pm\infty}\|e^{iH_{\text{FW}}t}e^{-iH_{\text{FW}}^0t}U_0\Psi-U\Phi_\pm\| \tag{6.17b}$$

$$+\lim_{t\to\pm\infty}\|(UU_0^*-1)e^{-iH_{\text{FW}}^0t}U_0\Psi\|\,, \tag{6.17c}$$

where (cf. Sec. III)

$$U=a_+(H_{\text{FW}})+ia_-(H_{\text{FW}})\,\text{sgn}\,Q'\,, \tag{6.18a}$$

$$a_\pm(H_{\text{FW}})=(1/\sqrt2)(1\pm m|H_{\text{FW}}|^{-1})^{1/2} \tag{6.18b}$$

(U_0 defined in the same way with Q_0',H_{FW}^0).

From existence of (6.16) we conclude that (6.17b) vanishes, if

$$\Phi_\pm=U^*\Omega_\pm(H_{\text{FW}},H_{\text{FW}}^0)U_0\Psi\,. \tag{6.19}$$

From the intertwining relations (Ref. 24, p. 17) we obtain for $\Psi\in\mathscr{H}_{\text{a.c.}}(H_0)$

$$\lim_{|t|\to\infty}\|\{f(H_{\text{FW}})-f(H_{\text{FW}}^0)\}e^{-iH_{\text{FW}}^0t}\Psi\|=0 \tag{6.20}$$

for any bounded, continuous function f. Similarly we obtain for $\Psi\in\mathscr{D}$

$$\lim_{|t|\to\infty}\|\{|Q|-|Q_0|\}e^{-iH_{\text{FW}}^0t}\Psi\|=0 \tag{6.21}$$

since $|Q|=(H_{\text{FW}}^2-m^2)^{1/2}=H_{\text{FW}}\cdot f(H_{\text{FW}})$, where f is bounded. Now we can estimate (6.17c):

$$\|(U-U_0)e^{-iH_{\text{FW}}^0t}\chi\|$$

$$\leqslant\|\{a_+(H_{\text{FW}})-a_+(H_{\text{FW}}^0)\}e^{-iH_{\text{FW}}^0t}\chi\| \tag{6.22a}$$

$$+\|\{a_-(H_{\text{FW}})-a_-(H_{\text{FW}}^0)\}e^{-iH_{\text{FW}}^0t}\chi\| \tag{6.22b}$$

$$+\|(\text{sgn}\,Q'-\text{sgn}\,Q_0')e^{-iH_{\text{FW}}^0t}a_-(H_{\text{FW}}^0)\chi\|\,, \tag{6.22c}$$

where (6.22a) and (6.22b) vanish, as $|t|\to\infty$, because of (6.20). In view of (6.21), vanishing of (6.22c) is implied by (6.14) for χ in the set $\{a_-^{-1}|Q_0|\Psi|\Psi\in\mathscr{D}\}$, which is dense in $F(a<Q_0^2<b)\mathscr{H}_{\text{a.c.}}(Q_0^2)$. Just note that

$$\|(\text{sgn}\,Q'-\text{sgn}\,Q_0')e^{-iH_{\text{FW}}^0t}|Q_0|\Psi\|$$

$$\leqslant\|(Q-Q_0)e^{-iH_{\text{FW}}^0t}\Psi\|+\|(|Q|-|Q_0|)e^{-iH_{\text{FW}}^0t}\Psi\|\,.$$

This completes the proof of existence of (6.3).

From

$$(\text{sgn}\,Q')H_{\text{FW}}=-H_{\text{FW}}\,\text{sgn}\,Q' \tag{6.23}$$

we obtain

$$(\text{sgn}\,Q')\Omega_\pm(H_{\text{FW}},H_{\text{FW}}^0)=\Omega_\mp(H_{\text{FW}},H_{\text{FW}}^0)\,\text{sgn}\,Q'$$

and therefore

$$\begin{aligned}\Omega_\pm(H,H_0)&=U^*\Omega_\pm(H_{\text{FW}},H_{\text{FW}}^0)U_0\\&=(a_+-ia_-\,\text{sgn}\,Q')\\&\quad\times\Omega_\pm(a_+^0+ia_-^0\,\text{sgn}\,Q_0')\\&=\Omega_\mp(a_-^{02}-ia_+^0a_-^0\,\text{sgn}\,Q_0')\\&\quad+\Omega_\mp(a_-^{02}-ia_+^0a_-^0\,\text{sgn}\,Q_0')\\&=\tfrac12\Omega_\pm(1+\tau H_0|H_0|^{-1})\\&\quad+\tfrac12\Omega_\mp(1-\tau H_0|H_0|^{-1})\,,\end{aligned}$$

which together with (6.16) and

$$F(H_0 \gtrless 0) = \tfrac{1}{2}(1 \pm H_0 |H_0|^{-1})$$

finally gives the result (6.15). □

In the example above, $Q - Q_0 = \boldsymbol{\alpha}\cdot\mathbf{A}(\mathbf{x})$. Condition (6.14) becomes

$$\lim_{|t|\to\infty} \|\mathbf{A}(\mathbf{x})\exp\{-i\sqrt{-\Delta + m^2 t}\}\Psi\| = 0,$$

which for Ψ having a Fourier transform in $C_0^\infty(\mathbb{R}^3 \setminus \{0\})$ follows almost trivially from a nonstationary phase argument.

APPENDIX: NORMAL FORMS IN SPECIAL SITUATIONS

Here we list the operators $Q_\pm^2$ in the special situations of Sec. II. Together with (3.26) and (3.27) this gives the explicit diagonal (resp. off-diagonal) form of the corresponding concrete Dirac operator,

$$
\begin{aligned}
&(2.1)\colon\ Q_\pm^2 = p^2,\\
&(2.8)\colon\ Q_\pm^2 = (\mathbf{p}-\mathbf{A})^2 - \boldsymbol{\sigma}\cdot\mathbf{B},\\
&(2.9)\colon\ Q_\pm^2 = (\mathbf{p}-\mathbf{A})^2 + \delta^2 B^2 \mp \delta \operatorname{div}\mathbf{B}\\
&\qquad\qquad - \boldsymbol{\sigma}\cdot\mathbf{B} \mp 2\delta\boldsymbol{\sigma}\cdot(\mathbf{A}\wedge\mathbf{B})\\
&\qquad\qquad - 2\delta\boldsymbol{\sigma}\cdot(\mathbf{B}\wedge\mathbf{p}) \mp i\delta\boldsymbol{\sigma}\cdot(\operatorname{rot}\mathbf{B}),\\
&(2.10)\colon\ Q_\pm^2 = p^2 + \mu^2 E^2 \mp \mu \operatorname{div}\mathbf{E}\\
&\qquad\qquad - 2\mu\boldsymbol{\sigma}\cdot(\mathbf{E}\wedge\mathbf{p}) \mp i\mu\boldsymbol{\sigma}\cdot(\operatorname{rot}\mathbf{E}),\\
&(2.11)\colon\ Q_\pm^2 = -\frac{d^2}{dr^2} + \frac{\kappa(\kappa\mp 1)}{r^2}\\
&\qquad\qquad - \mu\left(\frac{2\kappa E(r)}{r} \mp \frac{dE(r)}{dr}\right)\\
&\qquad\qquad + \mu^2 E^2(r),\\
&(2.13)\colon\ Q_\pm^2 = p^2 \pm \boldsymbol{\sigma}\cdot\operatorname{grad} V(x) + (m + V(x))^2.
\end{aligned}
$$

Since the domain questions are more delicate in the Klein–Gordon case we add a few remarks: H, as defined in (2.16) [cf. also (2.21), (2.22)] is self-adjoint on

$$\mathscr{D}(H) = \mathscr{D}(T) + \mathscr{D}(T^{1/2}) \subset \mathscr{H}_+ + \mathscr{H}_-.$$

The definition of $Q_\pm$ in (3.24) specializes to

$$
\begin{aligned}
Q_+^2 &= T \quad \text{on } \mathscr{D}(T^{3/2}) \subset \mathscr{H}_+,\\
Q_-^2 &= T \quad \text{on } \mathscr{D}(T) \subset \mathscr{H}_-
\end{aligned}
$$

[$\mathscr{D}(\cdot)$ always denotes the domain of the indicated operator in L^2, $\subset$ means the inclusion via the identity map]. Here H may be diagonalized by $\widetilde{U}$ defined in (5.14). In this case $S = T^{1/2}, S^* = T^{-1/2}$.

The form usually found in the literature[10,29] is obtained as follows: $\mathscr{H}_+$ is unitarily mapped onto $\mathscr{H}_-$ by the operator $T^{1/2}$. Therefore $\mathscr{H} = \mathscr{H}_+ + \mathscr{H}_-$ can be identified with $L^2(\mathbb{R}^3)^2$. The identification map

$$I = \begin{pmatrix} T^{1/2} & 0 \\ 0 & \mathrm{id} \end{pmatrix}\colon\ \mathscr{H} \to L^2(\mathbb{R}^3)^2$$

is a unitary isomorphism (id denotes the identity on $\mathscr{H}_- = L^2$). The operator

$$IU = 2^{-1/2}\begin{pmatrix} T^{1/2} & \mathbb{1} \\ -T^{1/2} & \mathrm{id} \end{pmatrix}$$

transforms H into

$$
\begin{aligned}
H_{\mathrm{diag}} &= (IU)H(IU)^*\\
&= \begin{pmatrix} T^{1/2} & 0 \\ 0 & -T^{1/2} \end{pmatrix}\\
&= \begin{pmatrix} ((p-A)^2 + m^2)^{1/2} & 0 \\ 0 & -((p-A)^2 + m^2)^{1/2} \end{pmatrix},
\end{aligned}
$$

which is defined in $L^2(R^3)^2$.

Remark: We may write $(p-A)^2 = \widehat{D}^*\widehat{D} = \widehat{D}\widehat{D}^*$ with

$$\widehat{D} = \left(\sum_{i=1}^{3} (p_i - A_i)^2\right)^{1/2} \equiv |p - A|$$

and apply to H_{diag} an inverse transformation $\widehat{U}^{-1}$ [defined as in (5.11) with $\widehat{D}$ and $M_\pm = m$] to obtain the following "Dirac form" of the Klein–Gordon equation:

$$i\frac{d}{dt}\Phi(t) = \begin{pmatrix} m & |p-A| \\ |p-A| & -m \end{pmatrix}\Phi(t),$$

with $\Phi(t) = \widehat{U}^{-1} IU \Psi(t)$.

[1] E. deVries, "Foldy–Wouthuysen transformations and related problems," Fortschr. Phys. **18**, 149 (1970).

[2] L. L. Foldy, "Synthesis of covariant particle equations," Phys. Rev. **102**, 568 (1956).

[3] L. L. Foldy and S. A. Wouthuysen, "On the Dirac theory of spin-½ particles and its nonrelativistic limit," Phys. Rev. **78**, 29 (1950).

[4] M. Cini and B. Touschek, "The relativistic limit of the theory of spin-½ particles," Nuovo Cimento **7**, 422 (1958).

[5] D. L. Weaver, "Exact diagonalization of relativistic Hamiltonians including a constant magnetic field," J. Math. Phys. **18**, 306 (1977).

[6] Tsai Wu-yang, "Energy eigenvalues for charged particles in a homogeneous magnetic field: An application of the Foldy–Wouthuysen transformation," J. Math. Phys. **7**, 1945 (1973).

[7] R. G. Osche, "Dirac and Dirac–Pauli equations in the Foldy–Wouthuysen representation," Phys. Rev. D **15**, 2181 (1976).

[8] N. V. Borisov, W. Müller, and R. Schrader, "Relative index theorems and supersymmetric scattering theory," Preprint FUB-HEP/86-7, Berlin, 1986.

[9] D. Bollé, F. Gesztesy, H. Grosse, W. Schweiger, and B. Simon, "Witten index, axial anomaly and Krein's spectral shift function in supersymmetric quantum mechanics," preprint, 1986.

[10] R. J. Cirincione and P. R. Chernoff, "Dirac and Klein–Gordon equations: Convergence of solutions in the nonrelativistic limit," Commun. Math. Phys. **79**, 33 (1981).

[11] P. A. Deift, "Applications of a commutation formula," Duke Math. J. **45**, 267 (1978).

[12] H. Grosse and L. Pittner, "Supersymmetric quantum mechanics defined as sesquilinear forms," Preprint UWthPh-1986-39, Wien, 1986.

[13] C. H. Wilcox, "Wave operators and asymptotic solutions of wave propagation problems of classical physics," Arch. Rat. Mech. Anal. **22**, 37 (1966).

[14] L. C. Biedenharn and L. P. Horwitz, "Chiral two-component spinors and the factorization of Kramer's equation," Found. Phys. **14**, 953 (1984).

[15] F. Gesztesy, H. Grosse, and B. Thaller, "First-order relativistic corrections and spectral concentration," Adv. Appl. Math **6**, 159 (1985); "Spectral concentration in the nonrelativistic limit," Phys. Lett. B **116**, 155 (1982).

[16] F. Gesztesy, H. Grosse, and B. Thaller, "A rigorous approach to relativistic corrections of bound state energies for spin-½ particles," Ann. Inst. H. Poincaré **40**, 159 (1984).

[17] F. Gesztesy, H. Grosse, and B. Thaller, "Efficient method for calculating

relativistic corrections for spin-$\frac{1}{2}$ particles," Phys. Rev. Lett. **50**, 625 (1983).

[18] F. Gesztesy, B. Simon, and B. Thaller, "On the self-adjointness of Dirac operators with anomalous magnetic moment," Proc. Am. Math. Soc. **94**, 115 (1985).

[19] H. Leinfelder and C. G. Simader, "Schrödinger operators with singular magnetic vector potentials," Math. Z. **176**, 1 (1981).

[20] L. C. Biedenharn, "The 'Sommerfeld Puzzle' revisited and resolved," Found. Phys. **13**, 13 (1983).

[21] C. V. Sukumar, "Supersymmetry and the Dirac equation for a central Coulomb field," J. Phys. A **18**, L697 (1985).

[22] J. Wu, A. Stahlhofen, L. C. Biedenharn, and F. Iachello, "A group theoretical calculation of the S-matrix for the Dirac Coulomb problem," preprint.

[23] T. Kato, *Perturbation Theory for Linear Operators* (Springer, Berlin, 1976), 2nd ed.

[24] M. Reed and B. Simon, *Methods of Modern Mathematical Physics III, Scattering Theory* (Academic, New York, 1979).

[25] M. Loss and B. Thaller, "Scattering of particles by long-range magnetic fields," Ann. Phys. (NY) **176**, 159 (1987).

[26] B. Thaller, "Relativistic scattering theory for long-range potentials of the nonelectrostatic type," Lett. Math. Phys. **12**, 15 (1986).

[27] M. Loss and B. Thaller, "Short-range scattering in long-range magnetic fields. The relativistic case," to be published in J. Diff. Eq.

[28] B. Thaller and V. Enss, "Asymptotic observables and Coulomb scattering for the Dirac equation," Ann. Inst. H. Poincaré **45**, 147 (1986).

[29] R. A. Weder, "Scattering theory for the Klein Gordon equation," J. Funct. Anal. **27**, 100 (1978).

Gauge theory of a group of diffeomorphisms. III. The fiber bundle description

Eric A. Lord
Department of Applied Mathematics, Indian Institute of Science, Bangalore 560 012, India
P. Goswami
Department of Physics, Indian Institute of Science, Bangalore 560 012, India

(Received 17 June 1986; accepted for publication 5 August 1987)

A new fiber bundle approach to the gauge theory of a group G that involves space-time symmetries as well as internal symmetries is presented. The ungauged group G is regarded as the group of left translations on a fiber bundle $G(G/H,H)$, where H is a closed subgroup and G/H is space-time. The Yang–Mills potential is the pullback of the Maurer–Cartan form and the Yang–Mills fields are zero. More general diffeomorphisms on the bundle space are then identified as the appropriate gauged generalizations of the left translations, and the Yang–Mills potential is identified as the pullback of the dual of a certain kind of vielbein on the group manifold. The Yang–Mills fields include a torsion on space-time.

I. INTRODUCTION

The exploitation of the structures known as fiber bundles,[1,2] for the formulation of Yang–Mills theories (gauge theories), has received a great deal of attention in recent years. In the conventional approach to the gauging of a symmetry group G, one puts a connection (Lie algebra valued one-form) on a principal fiber bundle $P(M,G)$ (M being space-time) and interprets a "gauge transformation" as a change of section in this bundle.[3–6] There are indications that this approach is not an appropriate one if G involves space-time symmetries. Consider, for example, Poincaré gauge theories,[7–10] affine gauge theories,[11–14] and conformal gauge theories.[15,16] The conventional fiber bundle descriptions of these theories employ *ad hoc* structures: second-order frames[16] (conformal) and affine frames[1–17] (Poincaré and affine).The translational gauge potentials on M and the tetrad on M turn out to be conceptually distinct entities, as has been pointed out by several authors[6,17]; this is a clear indication of the inappropriateness of the conventional fiber bundle description, in the case of space-time symmetries. In particular, the conventional fiber bundle description is not appropriate for a Poincaré gauge theory.

We present an alternative fiber bundle description of gauge theories that does not encounter the difficulties mentioned above, and that provides a unified scheme for describing space-time and internal symmetries, and their "gauged" generalizations.

Our approach is based on the properties of the fiber bundles $G(G/H,H)$, where G is a Lie group, H a Lie subgroup, and G/H is interpreted as space-time. The group to be gauged is the group G of left translations and a gauge transformation is a bundle automorphism. A connection will be defined essentially as a particular kind of vielbein on G.

The idea of formulating gauge theories on a principal fiber bundle $G(G/H,H)$, with G/H interpreted as space-time, is not new. It was proposed and investigated by Ne'eman and Regge,[18,19] who studied the problem of constructing Lagrangians on the bundle space and showed that the idea can be extended consistently to provide a framework for supergravity theories. Our scheme differs from theirs in several important respects; their concept of gauge transformation was different from ours and their connection vielbein was not specialized. Further investigations into the construction of Lagrangians in the scheme of Ne'eman and Regge have been made by Pérez-Rendon and Ruiperez.[20] In the present work we shall not consider the problem of constructing Lagrangian theories; our emphasis is on the geometrical structure only.

In recent work[21] we have shown how Poincaré gauge theory can be generalized to groups other than the Poincaré group (such as the affine, de Sitter, and conformal groups). The fiber bundle concept was not employed—the formalism dealt only with quantities defined as fields on space-time. The present work is a fiber bundle interpretation of these ideas.

II. THE PRINCIPAL FIBER BUNDLE *G(G/H,H)*

Let G be a Lie group and H a closed Lie subgroup. The orbits of the *right* action of H on G are the left cosets gH. They are the fibers of the principal fiber bundle $G(G/H,H)$ whose structural group is H (acting on the right). Denoting the general element of G by z, the *left translation* associated with an element of $g\in G$ is the diffeomorphism $z\rightarrow z'=gz$ on the group manifold. On account of the associative law, the left translations constitute a group of diffeomorphisms on G, isomorphic to G. Throughout most of this work, we impose none of the common restrictions on G (such as semisimplicity, compactness, and connectedness). We find that they are not necessary.

Let $\sigma\colon G/H\rightarrow G$ be a section on $G(G/H,H)$. (If no global section exists, σ can be a collection of local sections; the notation for dealing with this case becomes cumbersome but the principles we shall develop remain valid. For simplicity, we shall not enter into these details.) Then any element $z\in G$ can be uniquely expressed as a product

0022-2488/88/010258-10$02.50

$$z=\sigma(x)\chi\,, \tag{2.1}$$

with $x\in G/H, \chi\in H$ (the point $x\in G/H$ is just πz, where π is the canonical projection on the bundle).

Left translations map fibers to fibers and so induce, in an obvious way, diffeomorphisms on G/H. This enables us to associate, with any left translation $z'=gz$, an H-valued field $h(g,x)$ on G/H, defined by

$$g\sigma(x)=\sigma(x')h(g,x) \tag{2.2}$$

(where x' denotes the image of $x\in G/H$ under the diffeomorphism induced on G/H by $z\rightarrow z'=gz$). The geometrical interpretation of (2.2) is illustrated in Fig. 1, in which it is to be understood that g acts on the left and $h=h(g,x)$ on the right.

We shall use a prime to denote the transform of a tensor field under the action of a diffeomorphism. Thus the action of the diffeomorphism $z\rightarrow z'=f(z)$ on a p-form field $\phi=(f^{-1})^*\phi$ and the action on a vector field V will be written $V'=(df)V$. Let Ψ be a set of p-forms that transform linearly among themselves according to some matrix representation S of H, under the right action of H:

$$\Psi'=\overline{S}(h)\Psi \quad (z'=zh)\,. \tag{2.3}$$

In this definition Ψ is a pseudotensorial form of type $(\overline{S},\mathfrak{h})$, where $\mathfrak{h}$ is the Lie algebra of H. In the case of scalar fields, (zero-forms) this prescription is equivalent to the usual construction of an associated fiber bundle.[3] We shall refer to a condition of the form (2.3) as a *fiber condition*. A set of fields Ψ satisfying a fiber condition is determined on the whole of a fiber if its value at one point of the fiber is given. For a set Ψ of *scalar* fields, the fiber condition (2.3) can be written in the alternative form

$$\Psi(zh)=\overline{S}^{-1}(h)\Psi(z)\,. \tag{2.4}$$

In the case of a set Ψ of p-forms, a set ψ of p-forms on G/H can be defined as the pullback

$$\psi=\sigma^*\Psi\,. \tag{2.5}$$

For a set Ψ of *scalar* fields, this is simply

$$\psi(x)=\Psi(\sigma(x)) \tag{2.6}$$

and the fiber condition (2.3) then leads to the transformation law

$$\psi(x')=\overline{S}(h(g,x))\psi(x) \tag{2.7}$$

under the action of a *left translation* $z'=gz$.

Equations (2.2) and (2.7) are essentially the fundamental relations of the theory of nonlinear realizations.[22–24]

In the present interpretation, G/H is a space-time. Then (2.7) is the active transformation law of a set of physical fields under the action of both space-time symmetries and internal symmetries. The subgroup I whose left action does not affect the points of G/H is the maximal subgroup of H that is an invariant subgroup of G. The group of diffeomorphisms on G/H induced by the left action of G is isomorphic to G/I.[2] The group I can be interpreted as an internal symmetry group and G/I as a space-time symmetry group. The reasonable candidates for G/I are the Poincaré group (with G/H Minkowski space), the conformal group (with G/H Minkowski space, or, in the case of the "full" conformal group including inversions, Minkowski space augmented by a "light cone at infinity"[25]), the de Sitter group[26,27] (acting on de Sitter space G/H), and the affine group (with G/H a metricless four-space which can, however, in a dynamical theory, acquire a metric as a Goldstone field when the affine symmetry is broken[28,29]).

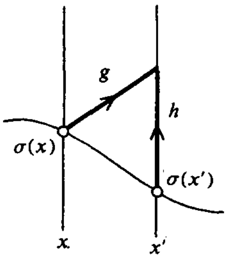

FIG. 1. The geometrical interpretation of (2.2).

III. DIFFERENTIAL GEOMETRY OF A GROUP MANIFOLD

In order to proceed further, we shall need to review briefly the elementary concepts from the differential geometry of a Lie group. The reader will almost certainly be familiar with these concepts, which are well known. However, our method of presentation is somewhat novel; it is aimed at establishing the results we shall need in the most rapid possible way.

Denote the general point of the manifold of a Lie group G by z. Denote an element of G regarded as an operator that acts on the group manifold (by left or right multiplication) by g. Of course, every element is simultaneously a point of the manifold and an operator on the manifold, but the conceptual distinction is, nevertheless, a useful one. Let e denote the unit element. Put a coordinate system on the group manifold and use the letters $M,N,\ldots$ for holonomic indices (so that the coordinates of z will be written z^M). Call this the "main" coordinate system. In addition, introduce an extra coordinate chart U, containing e, and use the letters $A,B,\ldots$ for coordinate labels in U.

In terms of coordinates, an infinitesimal left translation $z'=gz$ is given by $g^A=e^A-a^A$ (a^A infinitesimal), $z'^M=(gz)^M=z^M-a^A L_A{}^M$, where

$$L_A{}^M(z)=\left.\frac{\partial(gz)^M}{\partial g^A}\right|_{g=e}\,. \tag{3.1}$$

Write L for the matrix whose matrix elements are $L_A{}^M$ and write $L_M{}^A$ for the matrix elements of L^{-1}. The quantities $L_A{}^M$ are the components of a vielbein (linear frame) on the group manifold, which we shall call the *left vielbein*. The vector fields $L_A=L_A{}^M\partial_M$ ($\partial_M=\partial/\partial z^M$) provide a basis for the tangent space to G at each point, and the dual basis for the contangent space is provided by the one-forms $L^A=dz^M L_M{}^A$.

Under a change of the main coordinate system, $L'^M_A(z')=L_A{}^N(z)\partial z'^M/\partial z^N$. Under a change of the coordinatization on U, $L'_A{}^M=K_A{}^B L_B{}^M$, where K is a constant matrix ($K_A{}^B=\partial g^B/\partial g'^A|_{g=e}$). This corresponds to a change of basis of the Lie algebra (see below). The indices $A,B,\ldots$ have taken on the role of anholonomic indices.

In an exactly analagous manner, we can define

$$R_A{}^M(z) = \left.\frac{\partial(zg)^M}{\partial g^A}\right|_{g=e} . \tag{3.2}$$

Write R for the matrix whose matrix elements are $R_A{}^M$ and write $R_M{}^A$ for the matrix elements of R^{-1}. We have a *right vielbein* consisting of the vector fields $R_A = R_A{}^M \partial_M$ and the dually related one-forms $R^A = dz^M R_M{}^A$.

A left-invariant vector field X is one for which $X' = X$ under any left translation. The left-invariant vector fields are the linear combinations of the vector fields R_A, with constant coefficients. In particular, the vector fields R_A are left invariant, so $R_A{}^M(z') = R_A{}^N(z)\partial z'^M/\partial z^N$ $(z' = gz)$. Therefore, for a left translation,

$$\frac{\partial z'^M}{\partial z^N} = [R^{-1}(z)R(z')]_N{}^M \quad (z' = gz) . \tag{3.3}$$

Similarly, right-invariant vector fields are linear combinations of L_A, and, for a right translation,

$$\frac{\partial z'^M}{\partial z^N} = [L^{-1}(z)L(z')]_N{}^M \quad (z' = zg) . \tag{3.4}$$

Now let S be a matrix field on G, providing a matrix representation of G. The *generators* of the representation are the matrices

$$G_A = \left.\frac{\partial S(g)}{\partial g^A}\right|_{g=e} . \tag{3.5}$$

Differentiating $S(gz) = S(g)S(z)$ with respect to g^A and setting $g = e$, we get

$$L_A S = G_A S . \tag{3.6}$$

Similarly,

$$R_A S = S G_A . \tag{3.7}$$

Hence

$$S^{-1} G_A S = D_A{}^B G_B , \tag{3.8}$$

where D is the matrix field

$$D = LR^{-1} . \tag{3.9}$$

It is obvious from (3.8) that the matrices D provide a representation of G. It is of course the *adjoint representation.* Write c_A for its generators and define the structure constants of G in terms of the matrix elements of the c_A:

$$c_{AB}{}^C = (c_B)_A{}^C . \tag{3.10}$$

Now write (3.8) in the form $G_A S(g) = D_A{}^C(g) S(g) G_C$, differentiate with respect to g^B and set $g = e$, and we obtain the familiar commutation relations

$$[G_A, G_B] = c_{AB}{}^C G_C . \tag{3.11}$$

From (3.6) and (3.7) we now readily derive the commutation relations satisfied by the left and right vielbeins:

$$[R_A, R_B] = c_{AB}{}^C R_C , \tag{3.12}$$

$$[L_A, L_B] = -c_{AB}{}^C L_C , \tag{3.13}$$

$$[R_A, L_B] = 0 . \tag{3.14}$$

These equations are the infinitesimal forms of the transformation laws of the left and right vielbeins under left and right translations. The finite forms are

$$L'_A = D_A{}^B(g^{-1}) L_B, \quad R'_A = R_A \quad (z' = gz) ; \tag{3.15}$$

$$L'_A = L_A, \quad R'_A = D_A{}^B(g) R_B \quad (z' = zg) . \tag{3.16}$$

(For example, under the left translation $z' = gz$, $L'_A{}^M(z') = L_A{}^N(z) \qquad [R^{-1}(z)R(z')]_N{}^M$, and $L(z)R^{-1}(z)R(z') = D(z)D^{-1}(z')L(z') = D^{-1}(g)L(z')$. This establishes the first of the above transformation laws. The others are proved similarly.)

The algebra of left-invariant vector fields, defined through the commutator product rule (3.12), is the *Lie algebra* of G. Because of (3.11), the generators of any matrix representation of G provide a matrix representation of the Lie algebra.

The $R_M{}^A$ are the components of the *Maurer–Cartan form*. This is the Lie-algebra valued one-form θ which, in any matrix representation of the Lie algebra with generators G_A, is represented by $R^A G_A$. From (3.7) we have $\partial_M S = R_M{}^A S G_A$ and therefore $R_M{}^A G_A = S^{-1} \partial_M S$. This can be written in a more abstract form, without reference to any particular representation, simply as

$$\theta = z^{-1} dz . \tag{3.17}$$

From (3.15) and (3.16) it follows that

$$\begin{aligned} R'^A &= R^A \quad (z' = gz) , \\ R'^A &= R^B D_B{}^A(g^{-1}) \quad (z' = zg) , \end{aligned} \tag{3.18}$$

which mean that the Maurer–Cartan form is left invariant and transforms under right translations according to the coadjoint representation $(D^T)^{-1}$. With the aid of (3.8) this behavior of the R^A under right translations can be formulated as $R'^A G_A = S(g) R^A G_A S(g^{-1})$, or, without reference to any particular representation, $\theta' = g\theta g^{-1}$. Thus we obtain the transformation laws of the Maurer–Cartan form under left and right translations:

$$\theta' = \theta \quad (z' = gz), \quad \theta' = g\theta g^{-1} \quad (z' = zg) . \tag{3.19}$$

[These laws can also be derived directly from the definition (3.17). We have $\theta'(z') = \theta(z)$ under any diffeomorphism. In particular, under $z' = gz$, $\theta'(z) = \theta(g^{-1}z)$ $= z^{-1} g d(g^{-1}z) = z^{-1} dz = \theta(z)$, and under $z' = zg$, $\theta'(z)$ $= \theta(zg^{-1}) = gz^{-1} d(zg^{-1}) = g\theta(z)g^{-1}$.]

In our notation, the *Maurer–Cartan equation*, in terms of components, is

$$\partial_M R_N{}^A - \partial_N R_M{}^A + R_M{}^B R_N{}^C c_{BC}{}^A = 0 . \tag{3.20}$$

It is equivalent to the commutator relation (3.12).

The (anholonomic) components of the *Cartan metric* on G are

$$\gamma_{AB} = -\operatorname{tr} c_A c_B = -c_{EA}{}^F c_{FB}{}^E . \tag{3.21}$$

They satisfy

$$\gamma_{AB} = D_A{}^C D_B{}^D \gamma_{CD} . \tag{3.22}$$

Proof: Apply (3.8) in the form $G_A S = D_A{}^B S G_B$ to the case where S is itself the adjoint representation. We get immediately

$$c_{AB}{}^E D_E{}^C = D_A{}^E D_B{}^F c_{EF}{}^C$$

and (3.22) follows. The quantity

$$c_{ABC} = c_{AB}{}^D \gamma_{DC} \tag{3.23}$$

is completely skew symmetric ($c_{ABC} = -c_{AB}{}^{D} \operatorname{tr} c_D c_C = -\operatorname{tr}[c_A c_B] c_C$, which is easily seen to be completely skew symmetric). The equation $c_{ABC} + c_{ACB} = 0$ is in fact the infinitesimal form of (3.22). The *holonomic* components of the Cartan metric are

$$\gamma_{MN} = L_M{}^A L_N{}^B \gamma_{AB} = R_M{}^A R_N{}^B \gamma_{AB} \,. \tag{3.24}$$

The two alternative expressions are identical on account of (3.22). Since the L^A are right invariant and the R^A left invariant, the holonomic metric is both left and right invariant. In other words, the left- and right-invariant vector fields on G are Killing fields for this metric.

IV. REFERENCE SYSTEM ON *G(G/H,H)*

Suppose that coordinate systems are given on G/H and on H. Let x^i denote the coordinates of a general point $x \in G/H$ and let χ^m denote the coordinates of a general point $\chi \in H$ (the letters $i,j,\ldots$ will be used throughout as holonomic indices for G/H, and the letters $m,n,\ldots$ will be used as holonomic indices for H). With reference to a chosen section σ, the prescription (2.1) induces a coordinatization of G, whereby z is given the set of coordinates

$$z^M = (x^i, \chi^m) \,. \tag{4.1}$$

The splitting of the holonomic indices

$$M = (i,m) \tag{4.2}$$

corresponds to the local homeomorphisms between $G(G/H,H)$ and $(G/H) \otimes H$. A similar splitting of the anholonomic indices,

$$A = (\alpha, a) \tag{4.3}$$

can also be introduced. Put an extra coordinate chart V on G/H, containing πe, and an extra coordinate chart W on H, containing $(\sigma \pi e)^{-1}$. We use the letters $\alpha,\beta\ldots$ as coordinate labels for V and the letters $a,b,\ldots$ as coordinate labels for W. The prescription (2.1) now determines a coordination of $U = \{z \in G: \pi z \in V, (\sigma \pi z)^{-1} \in W\}$, whereby $g \in U$ is assigned the set of coordinates

$$g^A = (g^\alpha, g^a) \,. \tag{4.4}$$

[It is possible to be slightly more general and to use different sections for establishing the coordinatization (4.1) of G and the coordinatization (4.4) of U.]

A reference system set up according to the above procedure is adapted to the fibration in a particularly useful way. We shall call such a reference system *canonical.* Transformations relating different canonical systems consist of coordinate changes in G/H and in H, and changes of section. Of particular importance among the possible changes of the canonical reference system are changes of basis of the Lie algebra

$$G'_A = K_A{}^B G_B \,, \tag{4.5}$$

where K has the special form

$$K_A{}^B = \begin{pmatrix} \delta^\beta_\alpha & K_\alpha{}^b \\ 0 & \delta_a{}^b \end{pmatrix} . \tag{4.6}$$

We shall refer to this as a *K transformation.*

By splitting g and zg according to the precription (2.1) it is not difficult to deduce from the definition (3.2) of the right vielbein that

$$R_a{}^i = 0 \tag{4.7}$$

in a canonical system. This implies (through $RR^{-1} = 1$) that

$$R_m{}^\alpha = 0 \,. \tag{4.8}$$

Thus in a canonical reference system the sets of components of the right vielbein and of the Maurer–Cartan form have the reduced forms

$$R_A{}^M = \begin{pmatrix} R_\alpha{}^i & R_\alpha{}^m \\ 0 & R_a{}^m \end{pmatrix}, \quad R_M{}^A = \begin{pmatrix} R_i{}^\alpha & R_i{}^a \\ 0 & R_m{}^a \end{pmatrix} \tag{4.9}$$

[where, of course, the matrix $(R_i{}^\alpha)$ is the inverse of $(R_\alpha{}^i)$, and the matrix $(R_m{}^a)$ is the inverse of $(R_a{}^m)$].

Equation (4.7) means that the vector fields R_a are tangential to the fibers. They are the "vertical" left-invariant vector fields.

The commutation relations (3.17) in conjunction with (4.7) imply that $c_{ab}{}^\gamma = 0$ in a canonical system $[c_{ab}{}^\gamma R_\gamma{}^i = c_{ab}{}^C R_C{}^i = R_a{}^M \partial_M R_b{}^i - R_b{}^M \partial_M R_a{}^i = 0]$. It will be useful to display the commutation relations (3.11) in the more specific form[21]

$$\begin{aligned} [G_\alpha, G_\beta] &= c_{\alpha\beta}{}^\gamma G_\gamma + c_{\alpha\beta}{}^c G_c \,, \\ [G_a, G_\beta] &= c_{a\beta}{}^\gamma G_\gamma + c_{a\beta}{}^c G_c \,, \\ [G_a, G_b] &= c_{ab}{}^c G_c \,. \end{aligned} \tag{4.10}$$

The $c_{ab}{}^c$ are just the structure constants of the subgroup H.

The vectors R_α span at each point of the bundle space G a "horizontal" subspace of the tangent space. Considering the Lie algebra as a vector space, the left-invariant vector fields R_α (represented by the G_α) span a subspace (not in general a Lie algebra) that we shall call the *translational part* of the Lie algebra of G. These concepts are not, in general, invariant concepts: horizontal vectors are not necessarily mapped to horizontal vectors under the (right) action of the structural group H of the bundle, and the translational part of the Lie algebra can be changed by a K transformation.

Now consider the adjoint representation of G, restricted to the subgroup H. On account of $c_{ab}{}^\gamma = 0$, we have

$$D_A{}^B(h) = \begin{pmatrix} D_\alpha{}^\beta(h) & D_\alpha{}^b(h) \\ 0 & D_a{}^b(h) \end{pmatrix} . \tag{4.11}$$

The matrices $(D_a{}^b(h))$ are the matrices of the adjoint representation of H and the matrices $(D_\alpha{}^\beta(h))$ provide a special matrix representation of H with the dimensionality of G/H. If $D_\alpha{}^b(h) = 0$ for all $h \in H$ or if all the $D_\alpha{}^b(h)$ can be transformed away by a K transformation, the space G/H is called *reductive*. In that case, there is a canonical system in which[21]

$$c_{\alpha c}{}^b = 0 \,. \tag{4.12}$$

The commutator $[R_\alpha, R_c]$ is then a linear combination of the R_β; the horizontal spaces are preserved by the (right) action of the structural group H and we have a "connection" in the Ehresmann sense on the fiber bundle $G(G/H,H)$. The components R^a of the Maurer–Cartan form of G are the components of the connection one-form of this Ehresmann connection. Thus the components R^a comprise a connection one-form on $G(G/H,H)$ if and only if G/H is reductive (see, for example, Kobayashi and Nomizu,[1] p. 103). In our scheme, we shall not insist that G/H be reductive. Indeed, the case when G/H is *not* reductive is especially interesting.

V. INFINITESIMAL GENERATORS AND COVARIANT DERIVATIVES OF FIELDS ON *G/H*

Let Ψ be a set of scalar fields on G, satisfying a fiber condition (2.3). We define the operators M_A and $\mathring{Q}_A$, which act on the pullback $\Psi = \sigma^*\Psi$ as follows:

$$M_A\psi = \sigma^*(L_A\Psi)\,, \tag{5.1}$$

$$\mathring{Q}_A\psi = \sigma^*(R_A\Psi)\,. \tag{5.2}$$

An infinitesimal left translation $z'^M = z^M - a^A L_A{}^M$ induces a diffeomorphism on G/H given by

$$x^i = x^i - a^A L_A{}^i\,. \tag{5.3}$$

[It is instructive to note that the fact that the quantities $L_A{}^i$ are functions of x only, $L_A{}^i = L_A{}^i(x)$ independent of χ, can be inferred from the commutation relations (3.14) together with (4.7); we have

$$\begin{aligned} R_a{}^m\,\partial_m L_A{}^i &= R_a{}^M\,\partial_M L_A{}^i - L_A{}^M\,\partial_M R_a{}^i \\ &= [R_a,L_A]^i = 0\,. \end{aligned}$$

So $\partial_m L_A{}^i = 0$.] Under the diffeomorphism (5.3), the transformation law of ψ is

$$\delta\psi = a^A M_A\psi\,. \tag{5.4}$$

This is the infinitesimal form of the nonlinear transformation law (2.7). The M_A are the generators of infinitesimal left translations, for the nonlinearity transforming ψ. Interpreting G/H as space-time, the transformation laws (5.3) and (5.4) give the action of a space-time symmetry, or combination of a space-time symmetry and an internal symmetry, on the points of space-time and on physical fields [for example, if G is SO(4,2) and H the 11-parameter subgroup corresponding to the subgroup of the conformal group that consists of Lorentz rotations, dilatations, and special conformal transformations, we will get the action of the conformal group on Minkowski space-time and on physical fields[30]].

Since the fields $L_A\Psi$ satisfy a fiber condition $(L_A\Psi)' = \overline{S}(h)L_A\Psi$ $(z' = zh)$ (a consequence of the right invariance of the L_A), the action of M_A on $M_B\psi$ is well defined. We have $M_AM_B\psi = \sigma^*(L_AL_B\Psi)$. Therefore the operators M_A satisfy the same commutation relations as the L_A:

$$[M_A,M_B] = -c_{AB}{}^C M_C\,. \tag{5.5}$$

Similarly, since the fields R_A satisfy a fiber condition

$$(R_A\Psi)' = D_A{}^B(h)\overline{S}(h)R_A\Psi \quad (z' = zh)$$

[a consequence of the transformation law of the R_A under the right action of H, given by (3.16)], we can deduce that $\mathring{Q}_A\mathring{Q}_B\Psi = \sigma^*(R_AR_B\Psi)$ and therefore that the operators $\mathring{Q}_A$ satisfy the same commutation relations as the R_A:

$$[\mathring{Q}_A,\mathring{Q}_B] = c_{AB}{}^C\mathring{Q}_C\,. \tag{5.6}$$

In a similar manner, we can deduce also that

$$[M_A,\mathring{Q}_B] = 0\,. \tag{5.7}$$

The fiber condition on $R_A\Psi$ leads immediately, as a consequence of the considerations of Sec. II, to the following transformation law of the fields $\mathring{Q}_A$ under a left translation:

$$(\mathring{Q}_A\psi)'(x') = D_A{}^B(h(g,x))\overline{S}(h(g,x))(\mathring{Q}_B\psi)(x) \qquad (z' = gz). \tag{5.8}$$

For reasons that will become apparent, we shall call $\mathring{Q}_A\psi$ a *covariant derivative* of ψ. [For the present, observe only that $\mathring{Q}_A\psi$ is constructed from the field ψ on G/H *and its derivatives* $\partial_i\psi$ and that $\mathring{Q}_A\psi$ transforms under the group G of left translations homogeneously, in spite of the fact that the transformation matrix $\overline{S}(h(g,x))$ is x dependent. These properties are characteristic of a covariant derivative.]

Finally, note the relation

$$M_A = D_A{}^B(\sigma)\mathring{Q}_B\,, \tag{5.9}$$

which is a direct consequence of the definitions of the operators M_A and $\mathring{Q}_A$.

VI. A CONNECTION ON *G/H*

We define the *connection* on G/H, associated with the group G of left translations, to be the pullback of the Maurer–Cartan form,

$$\mathring{\Gamma} = \sigma^*\theta\,. \tag{6.1}$$

Since $\pi\sigma = 1$, the components of $\sigma(x)$ have the form $\sigma^M(x) = (x^i,\sigma^m(x))$. The section is determined by the functions $\sigma^m(x)$. In terms of components, the definition (6.1) has the more explicit form

$$\mathring{\Gamma}_i{}^A(x) = \sigma^M{}_{.i}R_M{}^A(\sigma) = R_i{}^A(\sigma) + \sigma^m{}_iR_m{}^A(\sigma)\,. \tag{6.2}$$

In particular, the components of the translational part of the connection are

$$\mathring{e}_i{}^\alpha = \mathring{\Gamma}_i{}^\alpha = R_i{}^a(\sigma)\,. \tag{6.3}$$

The inverse $(\mathring{e}_\alpha{}^i)$ of the matrix $(\mathring{e}_i{}^\alpha)$ provides a set of components of a *vielbein* on G/H. Since we are interpreting G/H as a space-time, they are the components of a *tetrad*.

Let us define the matrices

$$\mathring{\Gamma}_i = \mathring{\Gamma}_i{}^AG_A = \mathring{e}_i{}^\alpha G_\alpha + \mathring{\Gamma}_i{}^aG_a\,, \tag{6.4}$$

where the G_A are the generators of any matrix representation S of G. In terms of the matrices (6.4), the definition (6.2) can be written as $\mathring{\Gamma}_i = \sigma^M{}_{.i}R_M{}^A(\sigma)G_A$. But from (3.7) we have $R_M{}^A(\sigma)G_A = S^{-1}(\sigma)S_{.M}(\sigma)$, so

$$\mathring{\Gamma}_i = \sigma^M{}_{.i}S^{-1}(\sigma)S_{.M}(\sigma) = S^{-1}(\sigma)\partial_iS(\sigma)\,.$$

So we can write, symbolically without reference to a particular representation,

$$\mathring{\Gamma}_i = \sigma^{-1}\,\partial_i\sigma\,. \tag{6.5}$$

Since the Maurer–Cartan form is left invariant, and since the section is not changed when a left translation is applied, the connection $\mathring{\Gamma}$ is left invariant,

$$\mathring{\Gamma}' = \mathring{\Gamma} \quad (z' = gz)\,. \tag{6.6}$$

Therefore

$$\mathring{\Gamma}_i{}'(x') = \mathring{\Gamma}_i(x') = (\sigma(x'))^{-1}\frac{\partial\sigma(x)}{\partial x'^i}$$

$$= \frac{\partial x^j}{\partial x'^i}(h\sigma^{-1}g^{-1})\partial_j(g\sigma h^{-1})$$

$$= \frac{\partial x^j}{\partial x'^i}(h(\sigma^{-1}\,\partial_j\sigma)h^{-1} + h\,\partial_j h^{-1}).$$

The transformation law of the components of $\mathring{\Gamma}$ under a left translation, under which they remain invariant, is therefore

$$\mathring{\Gamma}_i{}'(x') = \mathring{\Gamma}_i(x) = \frac{\partial x^j}{\partial x'^i}(h\mathring{\Gamma}_j h^{-1} - \partial_j h\cdot h^{-1}), \quad (6.7)$$

where

$$h = h(g,x) \quad (6.8)$$

is given by (2.2).

An alternative form of this transformation law is

$$\mathring{\Gamma}'_i{}^A(x') = \mathring{\Gamma}_i{}^A(x')$$

$$= \frac{\partial x^j}{\partial x'^i}(\mathring{\Gamma}_j{}^B(x)D_B{}^A(h^{-1})$$

$$+ \partial_j(h^{-1})^m\cdot R_m{}^A(h^{-1})). \quad (6.9)$$

That this is equivalent to (6.7) can be verified as follows. Multiply (6.9) by G_A and apply (3.8) to the first term. For the second term, we employ (3.7) in the form $R_M{}^{A\cdot}G_A = S^{-1}S_{\cdot M}$. This implies

$$\partial_j(h^{-1})^m R_m{}^A(h^{-1})G_A$$

$$= \partial_j(h^{-1})^m S(h)\partial_m S(h^{-1}) = S(h)\,\partial_j S(h^{-1}).$$

Equation (6.9) has now become

$$\mathring{\Gamma}_i{}'(x') = \mathring{\Gamma}_i(x')$$

$$= \frac{\partial x^j}{\partial x'^i}(S(h)\mathring{\Gamma}_j(x)S(h^{-1}) + S(h)\partial_j S(h^{-1})),$$

which is just (6.7) evaluated for a particular representation S of G. A direct derivation of (6.9) is rather more complicated. We have

$$\mathring{\Gamma}_i{}'^A(x') = \mathring{\Gamma}_i{}^A(x') = \frac{\partial\sigma^M(x')}{\partial x'^i}R_M{}^A(\sigma(x'))$$

$$= \frac{\partial x^j}{\partial x'^i}\frac{\partial(g\sigma h^{-1})^M}{\partial x^j}R_M{}^A(\sigma(x'))$$

and

$$\frac{\partial(g\sigma h^{-1})^M}{\partial x^j}R_M{}^A(\sigma(x'))$$

$$= \frac{\partial(\sigma h^{-1})^N}{\partial x^j}\frac{\partial(g\sigma h^{-1})^M}{\partial(\sigma h^{-1})^N}R_M{}^A(\sigma(x'))$$

$$= \left[\frac{\partial\sigma^P}{\partial x^j}\frac{\partial(\sigma h^{-1})^N}{\partial\sigma^P} + \frac{\partial(h^{-1})^m}{\partial x^j}\frac{\partial(\sigma h^{-1})^N}{\partial(h^{-1})^m}\right]$$

$$\times[R^{-1}(\sigma h^{-1})R(g\sigma h^{-1})]_N{}^M R_M{}^A(\sigma(x')).$$

The transformation law (6.9) implies the transformation law

$$\mathring{e}'_i{}^\alpha(x') = \mathring{e}_i{}^\alpha(x') = \frac{\partial x^j}{\partial x'^i}\mathring{e}_j{}^\beta D_\beta{}^\alpha(h^{-1}) \quad (6.10)$$

for the tetrad components, under the action of a left translation.

The Maurer–Cartan equation implies a zero curvature for the connection on G/H,

$$\partial_i\mathring{\Gamma}_j{}^A - \partial_j\mathring{\Gamma}_i{}^A + \mathring{\Gamma}_i{}^B\mathring{\Gamma}_j{}^C c_{BC}{}^A = 0. \quad (6.11)$$

Proof: Let $R^{\,A}_{M\cdot N}(\sigma)$ denote $\partial_N R^{\,A}_M$ evaluated at $\sigma(x)$. Then

$$\partial_i\mathring{\Gamma}_j^A - \partial_j\mathring{\Gamma}_i^A = \partial_i(\sigma^M{}_{\cdot j}R^{\,A}_M(\sigma)) - \partial_j(\sigma^M{}_{\cdot i}R^{\,A}_M(\sigma))$$

$$= \sigma^M{}_{\cdot j}\sigma^N{}_{\cdot i}(R^{\,A}_{M\cdot N}(\sigma) - R^{\,A}_{N\cdot M}(\sigma))$$

$$= \sigma^M{}_{\cdot i}\sigma^N{}_{\cdot j}R^{\,B}_M(\sigma)R^{\,C}_N(\sigma)c_{BC}{}^A$$

$$= \mathring{\Gamma}_j{}^B\mathring{\Gamma}_i{}^C c_{BC}{}^A.$$

We are now in a position to obtain a more explicit expression for the covariant derivative $\mathring{Q}_A\psi$ of a field on G/H. We have

$$\mathring{Q}_A\psi = \sigma^*(R_A\Psi) = R_A{}^M(\sigma)\Psi_{\cdot M}(\sigma), \quad (6.12)$$

where $\Psi_{\cdot M}(\sigma)$ denotes $\partial_M\Psi$ evaluated at the point $\sigma(x)$, so that

$$\partial_i\psi = \partial_i\Psi(\sigma) = (\partial_i\sigma^M)\Psi_{\cdot M}(\sigma)$$

$$= \Psi_{\cdot i}(\sigma) + (\partial_i\sigma^m)\Psi_{\cdot m}(\sigma). \quad (6.13)$$

Differentiate the fiber condition $\Psi(z) = \overline{S}(h)\Psi(zh)$ with respect to h^a and set $h = e$. We got $0 = \overline{G}_a\Psi + R_a\Psi = \overline{G}_a\Psi + R_a{}^M\,\partial_M\Psi$ (where the $\overline{G}_a$ are the generators of the matrix representation $\overline{S}$ of H). Hence

$$\Psi_{\cdot m}(\sigma) = -R_m{}^a(\sigma)\overline{G}_a\psi. \quad (6.14)$$

By substituting (6.13) and (6.14) into (6.12), we find

$$\mathring{Q}_\alpha\psi = \mathring{D}_\alpha\psi = \mathring{e}_\alpha{}^i\mathring{D}_i\psi, \quad \mathring{D}_i\psi = \partial_i\psi + \mathring{\Gamma}_i{}^a\overline{G}_a\psi, \quad (6.15)$$

and

$$\mathring{Q}_a\psi = -\overline{G}_a\psi. \quad (6.16)$$

Observe that, although $\mathring{D}_i$ *looks* like a convariant derivative associated with the group H, in general it is not. This phenomenon was encountered already in our earlier work.[21] The present fiber bundle description gives a much clearer geometrical insight into what is happening. The $\mathring{Q}_\alpha\psi$ and $\mathring{Q}_a\psi$ are two pieces of a single geometrical entity, which transforms according to the transformation law (5.8). *Only when G/H is reductive* do these two pieces transform independently and in that case $D_i\psi$ transforms like a true covariant derivative for the subgroup H of the group G of left translations.

VII. METRICS INDUCED ON *G/H*

In certain circumstances the diffeomorphisms induced on G/H by left translations on G are closely related to naturally arising metrical properties of G/H.

The most straightforward case arises when G is semisimple, so that the Cartan metric γ_{AB} has an inverse γ^{AB}, and *in addition* the submatrix $\gamma^{\alpha\beta}$ is nonsingular (we shall denote its inverse by $\eta_{\alpha\beta}$). We can in that case define a nonsingular rank 2 tensor field on G/H,

$$g^{ij} = \gamma^{AB}L_A{}^iL_B{}^j = \gamma^{AB}R_A{}^iR_B{}^j = \gamma^{\alpha\beta}R_\alpha{}^iR_\beta{}^j = \gamma^{\alpha\beta}\mathring{e}_\alpha{}^i\mathring{e}_\beta{}^j. \quad (7.1)$$

Since the quantities $L_A{}^i$ depend only on x, the g_{ij} are indeed uniquely defined on G/H. The fact that the vectors R_A are left invariant ensures that the g^{ij} are invariant under the diffeomorphisms induced on G/H by the left translations. These diffeomorphisms are therefore *isometries* for the metric on G/H with components $g_{ij} = \mathring{e}_i{}^\alpha \mathring{e}_j{}^\beta \eta_{\alpha\beta}$. The $L_A{}^i$ are the components of a set of Killing vectors on G/H. The transformation law of the metric of G/H under left translations (under which the components remain invariant) is

$$g_{ij}(x') = g'_{ij}(x') = \frac{\partial x^k}{\partial x'^i} \frac{\partial x^l}{\partial x'^j} g_{kl}(x) . \tag{7.2}$$

An example of an induced metric of this kind is the metric of de Sitter space G/H, where G is SO(4,1).

If the conditions on G that lead to the above construction of a metric on G/H are not satisfied, there may nevertheless exist a constant nonsingular matrix $\eta_{\alpha\beta}$ such that[21]

$$D_\alpha{}^\gamma(h) D_\beta{}^\delta(h) \eta_{\gamma\delta} = \rho(h) \eta_{\alpha\beta} \tag{7.3}$$

for every $h \in H$ (ρ is therefore a one-dimensional representation of H). In that case we can regard

$$g_{ij} = \mathring{e}_i{}^\alpha \mathring{e}_j{}^\beta \eta_{\alpha\beta} \tag{7.4}$$

as the components of a metric on G/H. It is left invariant because the tetrad is; under left translations $g'_{ij} = g_{ij}$. But the diffeomorphisms on G/H induced by left translations are not, *in general*, isometries because the transformation law of g_{ij} under left translation is not the usual tensor transformation law (7.2). The transformation law (6.10) of the tetrad gives

$$g'_{ij}(x') = g_{ij}(x') = \rho^{-1}(h(g,x)) \frac{\partial x^k}{\partial x'^i} \frac{\partial x^l}{\partial x'^j} g_{kl}(x) . \tag{7.5}$$

In the simplest case, $\rho = 1$ and the left translations induce isometries. An example of this is the action of the Poincaré group on Minkowski space. More generally, the left translations induce *conformal* mappings on G/H. Examples are the action of the conformal group SO(4,2) on Minkowski space and the action of the conformal group SO(4,2) on de Sitter space. These two cases are related by a K transformation.

Finally, there are cases in which a metric cannot be induced on G/H by the above method because no $\eta_{\alpha\beta}$ with the required property exists. An example of this is the action of the affine group on space-time.

VIII. GAUGING THE LEFT TRANSLATIONS

Up to this point, we have dealt only with the formalism associated with the *ungauged* group G of left translations. We shall now introduce those diffeomorphisms on G that can be regarded as the gauged generalizations of left translations.

Observe first that the (ungauged) left translations are just those diffeomorphisms $z \to z'$ on G that satisfy

$$z' g_0 = (z g_0)', \quad \text{for all } g_0 \in G . \tag{8.1}$$

We now define a *gauge transformation* to be a bundle automorphism, that is, a diffeomorphism $z \to z'$ on G that satisfies

$$z' h = (zh)', \quad \text{for all } h \in H . \tag{8.2}$$

(This concept of gauge transformation appears in the work of Atiyah, Hitchin, and Singer,[31] but we have abandoned the requirement that the action induced on base space shall be trivial.)

It then follows that the gauge transformations are just those diffeomorphisms on G of the form

$$z' = g(x) z , \tag{8.3}$$

where $x = \pi z$ and $g(x)$ is a G-valued function on G/H. [As proof, define $g(z) = z' z^{-1}$. Then (8.3) implies $g(z) = g(zh)$, for all $h \in H$. So $g(z)$ is constant on each fiber. We have a G-valued function $g(x) = g(z)$ on G/H.]

Observe that the form (8.3) of a gauge transformation is in agreement with the elementary concept of gauging a symmetry group, due to Yang and Mills; the group action is generalized by allowing the group elements to be space-time dependent.

It is important to note that not every G-valued field $g(x)$ on G/H defines a gauge transformation through (8.3). In general, a mapping on G of the form (8.3) will not even be one-to-one.

The geometrical meaning of (8.2) is that the gauge transformations are those diffeomorphisms on G that map fibers to fibers and preserve the action of the structural group H of $G(G/H,H)$. That is, two points on the same fiber related to each other by right multiplication by $h \in H$ will have two images related in the same way.

The prescription (2.2) generalizes immediately to the case of a gauge transformation. Simply replace g by $g(x)$ in (2.2). Figure 1 now illustrates this more general situation. In other words, a gauge transformation induces a diffeomorphism $x \to x'$ on G/H and specifies a unique H-valued field $h(x)$ on G/H, through the prescription[32]

$$h(x) = (\sigma(x'))^{-1} g(x) \sigma(x) . \tag{8.4}$$

Conversely, any diffeomorphism $x \to x'$ on G/H together with any H-valued field $h(x)$ on G/H determine a unique gauge transformation $z' = g(x) z$, through the prescription

$$g(x) = \sigma(x') h(x) (\sigma(x))^{-1} . \tag{8.5}$$

In Sec. II we obtained the transformation law of a field ψ on G/H, defined as the pullback of a scalar field Ψ on G satisfying a fiber condition. This transformation law generalizes immediately: the transformation law of ψ under a gauge transformation is simply

$$\psi'(x') = \overline{S}(h(x)) \psi(x) . \tag{8.6}$$

Consider now the *infinitesimal* gauge transformations. We have already seen that an infinitesimal (ungauged) left translation $z'^M = z^M - \Lambda^M$ is generated by an infinitesimal vector Λ that is *right invariant*,

$$[R_A, \Lambda] = 0 , \tag{8.7}$$

or, equivalently,

$$\Lambda = a^A L_A \quad (a^A \text{ const}) . \tag{8.8}$$

An infinitesimal gauge transformation is generated by an infinitesimal vector Λ that is *invariant under the right action of the subgroup H*,

$$[R_a, \Lambda] = 0 , \tag{8.9}$$

or, equivalently,

$$\Lambda = a^A L_A \quad (a^A = a^A(x)) . \tag{8.10}$$

The effect of an infinitesimal gauge transformation on the points of G/H and on the fields ψ on G/H is obtained as follows. Define $\lambda^M = \Lambda^M(\sigma)$. Then the diffeomorphism $x \to x'$ induced on G/H by an infinitesimal gauge transformation $z'^M = z^M - \Lambda^M = z^M - a^A(x)L_A{}^M$ is $x'^i = x^i - \lambda^i$ (note that Λ^i is dependent only on x anyway, so in fact $\lambda^i = \Lambda^i$). The transformation law of a scalar field Ψ on G is $\delta\Psi = \Lambda^M \partial_M \Psi$ so the infinitesimal form of (8.6) is $\delta\psi = \sigma^*(\Lambda^M \partial_M \Psi) = a^A M_A \psi$. Alternatively, regarding the λ^M as the (space-time-dependent) parameters rather than the a^A, $\delta\psi = \lambda^M \Psi_{\cdot M}(\sigma)$. Employing formulas (6.13) and (6.14) we find

$$\delta\psi = \lambda^i \partial_i \psi - \epsilon^a \overline{G}_a \psi, \tag{8.11}$$

where

$$\epsilon^a = (\lambda^M - \lambda^i \sigma^M{}_{\cdot i}) R_M{}^a(\sigma) = (\lambda^m - \lambda^i \sigma^m{}_{\cdot i}) R_m{}^a(\sigma). \tag{8.12}$$

The form (8.11) of the transformation law of ψ emphasizes the fact that a gauge transformation consists of a general diffeomorphism on the space-time G/H together with a space-time-dependent element of H. The geometrical explanation of the peculiar form of the parameters ϵ^a is given by Fig. 2; the vector λ can be built up from a component tangential to the section and a vertical component ϵ. The ϵ^a are the anholonomic components of ϵ.

IX. DIFFERENTIAL GEOMETRY OF A VIELBEIN

As a preliminary to the construction of a connection and curvature associated with our gauge transformations, we shall consider the differential geometry of a vielbein field on a differentiable manifold.

Denote the coordinates of the general point of a manifold by z^M (M,N holonomic indices; $A,B,\ldots$ anholonomic indices). Let E be the matrix of components $E_A{}^M$ of a vielbein and denote the matrix elements of E^{-1} by $E_M{}^A$. The vielbein vector fields are $E_A = E_A{}^M \partial_M$ and the dually associated one-form fields are $E^A = dz^M E_M{}^A$. The components $\Omega_{AB}{}^C$ of the "object of anholonimity" are defined by

$$[E_A, E_B] = \Omega_{AB}{}^C E_C . \tag{9.1}$$

Using the vielbein components to convert anholonomic indices to holonomic indices and vice versa, in the usual way, we have

$$\partial_M E_N{}^A - \partial_N E_M{}^A = -\Omega_{MN}{}^A . \tag{9.2}$$

Under the action of an infinitesimal diffeomorphism $z'^M = z^M - \Lambda^M$, the transformation law of the vielbein is

$$\delta E_A = [\Lambda, E_A], \tag{9.3}$$

which leads to

$$\delta E_M{}^A = \mathscr{D}_M \Lambda^A \tag{9.4}$$

where

$$\mathscr{D}_M \Lambda^A = \partial_M \Lambda^A + \Lambda^B \Omega_{MB}{}^A . \tag{9.5}$$

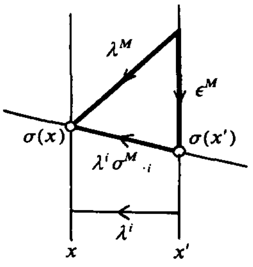

FIG. 2. The geometrical explanation of the form of the parameters ϵ^a.

Observe incidentally that the $\Omega_{MB}{}^A$ are the anholonomic components of the linear connection whose holonomic components are

$$\Gamma_{MN}{}^P = (\partial_N E_M{}^A) E_A{}^P = -E_M{}^A \partial_N E_A{}^P . \tag{9.6}$$

The operator $\mathscr{D}_M$ is then a covariant derivative operator. We have, for example, for a contravariant vector A^P,

$$\mathscr{D}_M A^P = \partial_M A^P + \Lambda^N \Gamma_{MN}{}^P . \tag{9.7}$$

The linear connection (9.5) has vanishing curvature but nonvanishing torsion $\Omega_{MN}{}^P$. With respect to this connection the manifold is a "space of distant parallelism."

When the manifold is the manifold of a Lie group G, the fields $G_{AB}{}^C = \Omega_{AB}{}^C - c_{AB}{}^C$ are of particular importance. We have

$$[E_A, E_B] - c_{AB}{}^C E_C = G_{AB}{}^C E_C \tag{9.8}$$

or equivalently,

$$\partial_M E_N{}^A - \partial_N E_M{}^A + E_M{}^B E_N{}^C c_{BC}{}^A = -G_{MN}{}^A . \tag{9.9}$$

Note that $G_{MM}{}^A$ vanishes for the right vielbein. The transformation law (9.4) can be rewritten in the form

$$\delta E_M{}^A = \Lambda^N G_{MN}{}^A + \nabla_M \Lambda^A , \tag{9.10}$$

where

$$\nabla_M \Lambda^A = \partial_M \Lambda^A - \Lambda^B E_M{}^C c_{BC}{}^A . \tag{9.11}$$

X. CONNECTION AND CURVATURE FOR THE GAUGE TRANSFORMATIONS

We shall now show how the formalism of our earlier work[21] on the gauging of a group G of space-time and internal symmetries arises as a particular case of the fiber bundle geometry.[33]

Define a *connection* on $G(G/H,H)$ to be the set of one-forms E^A dual to a vielbein E_A on the group manifold G, satisfying the following two conditions.

(1) The vielbein satisfies the same fiber condition as the right vielbein, namely

$$E'_A = D_A{}^B(h) E_B \quad (z' = zh), \tag{10.1}$$

for every $h \in H$ [see (3.16)].

(2) The vertical vectors of the vielbein are the same as those of the right vielbein ($E_a{}^M = R_a{}^M$). Thus in a canonical reference system the components of the specialized vielbein and its dually associated connection are

$$E_A{}^M = \begin{pmatrix} E_\alpha{}^i & E_\alpha{}^m \\ 0 & R_a{}^m \end{pmatrix}, \quad E_M{}^A = \begin{pmatrix} E_i{}^\alpha & E_i{}^a \\ 0 & R_m{}^a \end{pmatrix}. \tag{10.2}$$

The above two conditions are gauge invariant. The gauge invariance of (1) follows simply from the fact that gauge transformations, by definition, commute with the right action of H. Condition (2), $E_a = R_a$, transforms to $E'_a = R'_a$ (under a gauge transformation $z \to z'$). But by (8.9), the R_a are invariant, $R'_a = R_a$.

We call the set of two-forms whose components are $G_{MN}{}^A$ the *curvature* associated with the connection E^A. The infinitesimal form of the transformation law (10.1) is $[R_a, E_B] = c_{aB}{}^C E_C$. Since $R_a = E_a$, this implies that

${G_{aB}}^C = 0$. Therefore ${G_{mN}}^C = E^A_m E^B_N {G_{AB}}^C = 0$. Thus the only nonvanishing components of the curvature are ${G_{ij}}^A$.

We define the connection on G/H to be the pullback of the connection on G,

$$\Gamma^A = \sigma^* E^A . \tag{10.3}$$

Its components are

$$\Gamma_i{}^A = \sigma^M{}_{\cdot i} E_M{}^A(\sigma) . \tag{10.4}$$

In particular,

$$e_i{}^\alpha = \Gamma_i{}^\alpha = E_i{}^\alpha(\sigma) \tag{10.5}$$

defines a tetrad on the space-time G/H. Because of the conditions (1) and (2), the $E_M{}^A$ are completely determined by the $\Gamma_i{}^A$.

We define the curvature on G/H to be the pullback of the curvature on G. Its components are

$$\sigma^M{}_{\cdot i}\sigma^N{}_{\cdot j} G_{MN}{}^A(\sigma) = G_{ij}{}^A(\sigma) . \tag{10.6}$$

Observe that

$$\begin{aligned}\partial_i \Gamma_j{}^A &= \partial_i(\sigma^M{}_{\cdot j} E_M{}^A(\sigma))\\ &= \sigma^M{}_{\cdot ji} E_M{}^A(\sigma) + \sigma^M{}_{\cdot j}\sigma^N{}_{\cdot i} E^A{}_{M\cdot N}(\sigma) .\end{aligned}$$

So

$$\begin{aligned}\partial_i \Gamma_j{}^A - \partial_j \Gamma_i{}^A &= \sigma^M{}_{\cdot j}\sigma^N{}_{\cdot i}(E^A_{M\cdot N}(\sigma) - E^A_{N\cdot M}(\sigma))\\ &= \sigma^M{}_{\cdot j}\sigma^N{}_{\cdot i}(G_{MN}{}^A(\sigma)\\ &\quad - E^B_M(\sigma)E^C_N(\sigma)c_{BC}{}^A) .\end{aligned}$$

Therefore the components of the curvature on G/H are given by

$$\partial_i \Gamma_j{}^A - \partial_j \Gamma_i{}^A + \Gamma_i{}^B \Gamma_j{}^C c_{BC}{}^A = -G_{ij}{}^A \tag{10.7}$$

[it is convenient to drop the argument, writing simply $G_{ij}{}^A$ to mean $G_{ij}{}^A(\sigma)$].

Denote the components of the pullback of the object of anholonimity by $F_{AB}{}^C$. The quantities $F_{ij}{}^\alpha$ and $F_{ij}{}^a$ defined by

$$F_{ij}{}^A = e_i{}^\beta e_j{}^\gamma F_{\beta\gamma}{}^A \tag{10.8}$$

are the components of the H torsion and the H curvature.[21]

The infinitesimal vector field Λ^M that generates an infinitesimal gauge transformation has anholonomic components $\Lambda^A = \Lambda^M E_M{}^A$. Since this vector field is invariant under the right action of H, it is completely determined by either $\lambda^M = \Lambda^M(\sigma)$ or by $\lambda^A = \Lambda^A(\sigma)$, either of which can be regarded as the set of (space-time-dependent) parameters of the infinitesimal gauge transformation. The transformation law for the components of the connection on G/H under an infinitesimal gauge transformation are now easily found. We have $\delta\Gamma_i{}^A = \sigma^M{}_{\cdot i}\delta E^{MA} = \sigma^M{}_{\cdot i}\Lambda^A{}_{\cdot M}(\sigma) - \lambda^B \Gamma_i{}^c F_{BC}{}^A$ (from 9.4). Therefore

$$\delta\Gamma_i{}^A = \mathscr{D}_i \lambda^A , \tag{10.9}$$

where

$$\mathscr{D}_i \lambda^A = \partial_i \lambda^A - \lambda^B \Gamma_i{}^C F_{BC}{}^A . \tag{10.10}$$

An alternative form of (10.9) is

$$\delta\Gamma_i{}^A = \nabla_i \lambda^A - \lambda^j G_{ji}{}^A, \tag{10.11}$$

where

$$\nabla_i \lambda^A = \partial_i \lambda^A - \lambda^B \Gamma_i{}^C c_{BC}{}^A \tag{10.12}$$

(note that $\lambda^i = \lambda^\alpha e_\alpha{}^i$). An alternative set of parameters for an infinitesimal gauge transformation consists of the λ^i and the components ϵ^a of the vertical vector introduced in (8.12) and Fig. 2. We have

$$\epsilon^A = (\lambda^M - \lambda^i \sigma^M{}_{\cdot i})E_M{}^A(\sigma) = \lambda^A - \lambda^i \Gamma_i{}^A \tag{10.13}$$

(which satisfies $\epsilon^\alpha = 0$). This relation appeared already in our earlier work. The fiber bundle formalism gives it a very clear geometrical meaning. In terms of these parameters,

$$\delta\Gamma_i{}^A = \lambda^j \partial_j \Gamma_i{}^A + (\partial_i \lambda^j)\Gamma_j{}^A + \nabla_i \epsilon^A . \tag{10.14}$$

This shows that $\Gamma_i{}^A$ transforms like a one-form under the diffeomorphism on G/H induced by a gauge transformation, and "like a Yang–Mills potential for an internal symmetry group G" under the action of the H-valued field on G/H associated with the gauge transformation. The transformation law under a *finite* gauge transformation is therefore a straightforward generalization of the transformation law given in Sec. VI for the "trivial" connection $\mathring{\Gamma}_i$ under left translations. A finite gauge transformation consists of a general diffeomorphism $x \to x'$ on G/H and an H-valued field $h = h(x)$ on G/H, according to (8.4). The transformation law of the components of the connection on G/H is

$$\begin{aligned}\Gamma'_i{}^A(x') = \frac{\partial x^j}{\partial x'^i}(&\Gamma_j{}^B(x)D_B{}^A(h^{-1})\\ &+ \partial_j(h^{-1})^m R_m{}^A(h^{-1})) ,\end{aligned} \tag{10.15}$$

or equivalently, for $\Gamma_i = \Gamma_i{}^A G_A = e_i{}^\alpha G_\alpha + \Gamma_i{}^a G_a$,

$$\Gamma'_i(x') = \frac{\partial x^j}{\partial x'^i}(h\Gamma_i(x)h^{-1} - \partial_i h \cdot h^{-1}) . \tag{10.16}$$

In particular, the tetrad transforms according to

$$e'_i{}^\alpha(x') = \frac{\partial x^j}{\partial x'^i} e_j{}^\beta(x) D_\beta{}^\alpha(h^{-1}) . \tag{10.17}$$

The transformation law of the curvature $G_{ij}{}^A$ on G/H is

$$G'_{ij}{}^A(x') = \frac{\partial x^k}{\partial x'^i}\frac{\partial x^l}{\partial x'^j} G_{kl}{}^B D_B{}^A(h^{-1}) \tag{10.18}$$

or equivalently, for $G_{ij} = G_{ij}{}^A G_A = G_{ij}{}^\alpha G_\alpha + G_{ij}{}^a G_a$,

$$G'_{ij}(x') = \frac{\partial x^k}{\partial x'^i}\frac{\partial x^l}{\partial x'^j} hG_{ij}(x)h^{-1} . \tag{10.19}$$

It can be shown that *if G/H is reductive*, the $F_{ij}{}^A$ transform like the $G_{ij}{}^A$. Otherwise, they have a complicated inhomogeneous transformation law. If the translational part of the Lie algebra is Abelian, then $F_{ij} = G_{ij}$.

In cases when a nonsingular matrix $\eta_{\alpha\beta}$ satisfying the conditions of Sec. VII exists, if follows from (10.17) that the transformation law of the G/H metric $g_{ij} = e_i{}^\alpha e_j{}^\beta \eta_{\alpha\beta}$ is

$$g'_{ij}(x') = \rho(h^{-1})\frac{\partial x^k}{\partial x'^i}\frac{\partial x^l}{\partial x'^j} g_{kl}(x) . \tag{10.20}$$

Thus, in general, in such cases gauge transformations induce Weyl (scale) transformations on the space-time metric, as well as diffeomorphisms.

XI. THE GENERALIZED COVARIANT DERIVATIVE

The covariant derivative operator associated with (ungauged) left translations introduced in Sec. V is readily generalized to a derivative operator covariant under gauge transformations. We simply define

$$Q_A \psi = \sigma^*(E_A \Psi) . \tag{11.1}$$

Since the E_A satisfy a fiber condition, the action of Q_A on $Q_B \psi$ is well defined, and in fact $Q_A Q_B \psi = \sigma^*(E_A E_B \Psi)$. Then (9.1) and (10.8) imply

$$[Q_A, Q_B] = F_{AB}{}^C Q_C , \tag{11.2}$$

or, more explicity,

$$\begin{aligned} [Q_\alpha, Q_\beta] &= F_{\alpha\beta}{}^\gamma Q_\gamma + F_{\alpha\beta}{}^c Q_c , \\ [Q_a, Q_\beta] &= c_{\alpha\beta}{}^\gamma Q_\gamma + c_{a\beta}{}^c Q_c , \\ [Q_a, Q_b] &= c_{ab}{}^c Q_c . \end{aligned} \tag{11.3}$$

In the particular case of Poincaré gauge theory, this "generalized Lie algebra" is already well known.[10] The arguments that led to (6.15) and (6.16) now provide the following explicit expressions for the generalized operators Q_A:

$$\begin{aligned} &Q_\alpha \psi = D_\alpha \psi = e_\alpha{}^i D_i \psi, \quad D_i \psi = \partial_i \psi + \Gamma_i{}^a \overline{G}_a \psi , \\ &Q_a \psi = -\overline{G}_a \psi . \end{aligned} \tag{11.4}$$

The covariant transformation law of $Q_A \psi$ under a gauge transformation is a straightforward generalization of (5.8),

$$(Q_A \psi)'(x') = D_A{}^B(h) \overline{S}(h) (Q_B \psi)(x) . \tag{11.5}$$

The transformation law (8.11) of ψ under an infinitesimal gauge transformation can be reexpressed in terms of the parameters λ^A. We have $\delta\psi = \sigma^* \delta\Psi = \sigma^*(\Lambda^A E_A \Psi) = \lambda^A \sigma^*(E_A \Psi)$. Therefore

$$\delta\psi = \lambda^A Q_A \psi = \lambda^\alpha D_\alpha \psi - \lambda^a \overline{G}_a \psi . \tag{11.6}$$

[1] S. Kobyashi and K. Nomizu, *Foundations of Differential Geometry* (Wiley, New York, 1963), Vol. I.

[2] N. Steenrod, *The Topology of Fibre Bundles* (Princeton U. P., Princeton, NJ, 1951).

[3] J. P. Harnad and R. B. Pettitt, J. Math. Phys. **17**, 1827 (1976).

[4] M. Daniel and C. M. Viallet, Rev. Mod. Phys. **52**, 175 (1980).

[5] T. Eguchi, P. B. Gilkey, and A. J. Hanson, Phys. Rep. **66**, 213 (1980).

[6] D. Ivanenko and G. Sardanashvily, Phys. Rep. **94**, 1 (1983).

[7] T. W. B. Kibble, J. Math. Phys. **2**, 212 (1961).

[8] N. Mukunda, in *Proceedings of the Workshop on Gravitation and Relativistic Astrophysics, Ahmedabad India*, edited by A. R. Prasanna, J. V. Narlikar, and C. V. Vishveshwara (Indian Academy of Sciences, Bangalore, India, 1982).

[9] P. von der Heyde, Phys. Lett. A **58**, 141 (1976).

[10] F. W. Hehl, in *Proceedings of the 6th Course of the International School of Cosmology and Gravitation*, edited by P. G. Bergmann and V. de Sabbata (Plenum, New York, 1978).

[11] E. A. Lord, Phys. Lett. A **65**, 1 (1978).

[12] F. W. Hehl, E. A. Lord, and Y. Ne'eman, Phys. Lett. B **71**, 432 (1977).

[13] F. W. Hehl, E. A. Lord, and Y. Ne'eman, Phys. Rev. D **17**, 428 (1978).

[14] Y. Ne'eman, in *To Fulfill a Vision: Jerusalem Einstein Contennial Symposium on Gauge Theories and Unification of Physical Forces*, edited by Y. Ne'eman (Addison–Wesley, Reading, MA, 1981).

[15] E. A. Lord and P. Goswami, Pramana **25**, 635 (1985).

[16] J. P. Harnad and R. B. Pettitt, in "Group theoretical methods in physics," *Proceedings of the Vth International Colloquium*, edited by R. T. Sharp and B. Kolman (Academic, New York, 1977).

[17] T. Dass, Pramana **23**, 433 (1984).

[18] Y. Ne'eman and T. Regge, Phys. Lett. B **74**, 54 (1978).

[19] Y. Ne'eman and T. Regge, Riv. Nuovo Cimento **1**, (1978).

[20] A. Pérez-Rendon and D. H. Ruiperez, in *Differential Geometric Methods in Mathematical Physics*, edited by S. Sternberg (Reidel, Dordrecht, 1984).

[21] E. A. Lord and P. Goswami, J. Math. Phys. **27**, 2415 (1986); E. A. Lord, *ibid.*, **27**, 3051 (1986).

[22] S. Coleman, J. Wess, and B. Zumino, Phys. Rev. **177**, 2239 (1969).

[23] C. G. Callan, S. Coleman, J. Wess, and B. Zumino, Phy. Rev. **177**, 2247 (1969).

[24] A. Salam and J. Strathdee, Phys. Rev. **184**, 1750, 1760 (1969).

[25] E. A. Lord, Int. J. Theor. Phys. **13**, 89 (1974).

[26] S. W. MacDowell and P. Mansouri, Phys. Rev. Lett. **38**, 739 (1977).

[27] A. A. Tseytlin, Phys. Rev. D **26**, 3327 (1982).

[28] E. A. Ivanov and V. I. Ogievetsky, Lett. Math. Phys. **1**, 309 (1976).

[29] Y. Ne'eman and D. Sijacki, Ann. Phys. (NY) **120**, 292 (1979).

[30] G. Mack and A. Salam, Ann. Phys. (NY) **53**, 174 (1969).

[31] M. F. Atiyah, N. J. Hitchin, and I. M. Singer, Proc. R. Soc. London Ser. A **362**, 425 (1978).

[32] It is to be understood that, when σ is a collection of local sections σ_α with supports $U_\alpha \in G/H$, U_α a covering of G/H, (8.4) becomes a prescription for a collection of H-valued functions $h_{\alpha\beta}$ defined by $h_{\alpha\beta}(x) = \sigma_\beta(x')^{-1} g(x) \sigma_\alpha(x)$, where $x \in U_\beta$ and $x' \in U_\alpha$. They have to satisfy $h_{\gamma\delta}(x) = \eta_{\delta\beta}(x') h_{\alpha\beta}(x) \eta_{\alpha\gamma}(x)$ $(x \in U_\gamma, x' \in U_\delta)$, where $\eta_{\alpha\gamma}(x)$ are the transition functions $\eta_{\alpha\gamma}(x) = [\sigma_\alpha(x)]^{-1} \sigma_\gamma(x)$.

[33] The sign conventions of Ref. 21 (which are well established in Poincaré gauge theory) differ from those that arise quite naturally in Secs. X and XI of the present paper. Essentially, the ϵ^A, λ^A, and $\Gamma_i{}^A$ of Ref. 21 are the $-\epsilon^A$, $-\lambda^A$, and $-\Gamma_i{}^A$ of the present work. The signs of a^A and M^A also differ.

Comment on the paper "An exactly soluble relativistic quantum two-fermion problem" [J. Math. Phys. 27, 3055 (1986)]

A. O. Barut[a)]
International Centre for Theoretical Physics, Trieste, Italy

R. A. Moore and T. C. Scott
Guelph-Waterloo Programme for Graduate Work in Physics, Waterloo Campus, University of Waterloo, Waterloo, Ontario, Canada, N2L 3G1

(Received 24 June 1987; accepted for publication 5 August 1987)

A comment is made on the separability of the center of mass and relative coordinates in the exact solution of a covariant two-body equation for two spin-$\frac{1}{2}$ particles.

I. INTRODUCTION

In a recent paper[1] a covariant two-body equation for two spin-$\frac{1}{2}$ particles was considered and an exact solution was presented in the center of mass system. Although this solution is correct, it was incorrectly stated at the beginning of the paper that the center of mass and relative coordinates are exactly separable. We wish to correct this statement and indicate the proper treatment of the equation in an arbitrary frame.

II. THEORY

With a choice of the spacelike surface perpendicular to $n_\mu = (1000)$, the covariant equation can be written in the Hamiltonian form as

$$\{(1/M)(m_1\boldsymbol{\alpha}_1 + m_2\boldsymbol{\alpha}_2)\cdot\mathbf{P} + [(\boldsymbol{\alpha}_1 - \boldsymbol{\alpha}_2)\cdot\mathbf{P} \\ - \beta_1 m_1 - \beta_2 m_2 - V(r)]\}\Phi = E\Phi, \qquad (1)$$

where **P** is the total momentum and **p** and **r** are the relative coordinates. Thus the Hamiltonian separates into a sum of two terms, one depending on the center of mass momentum, the other on the relative coordinates. However, the coefficients in these two terms depending on the spin matrices do not commute, hence the solution cannot be written as a product of two functions, one depending on the center of mass coordinates and one on the relative coordinates. Since **R** does not appear, we have always a factor $e^{i\mathbf{P}\cdot\mathbf{R}}$ so that we can treat **P** as a number in the momentum representation.

Equation (1) is a specific case of an infinite component wave equation generally written as

$$(J^\mu P_\mu - K)\Phi = 0,$$

or, with J_0 diagonal and equal to 1 as in our case,

$$(\mathbf{J}\cdot\mathbf{P} - K)\Phi = E\Phi, \qquad (2)$$

where **J** and **K** (which is a function of relative coordinates or internal degrees of freedom) do not commute. There is a general procedure to solve Eq. (2) in an arbitrary frame[2] and it was an oversight not to connect Eq. (1) with Eq. (2). The method consists in finding the appropriate boost operators **M**, solving the equation in the rest frame ($\mathbf{P} = 0$), and then boosting the result to an arbitrary frame: $\Phi(\mathbf{P}) = e^{i\boldsymbol{\xi}\cdot M}\Phi(0)$.

Another procedure in the present case is to separate the radial and angular part of Eq. (1) for a general **P**. Since in the momentum representation the coefficient of **P** is a finite matrix, the method of separation used in Ref. 3 can easily be extended.

[a)] Permanent address: Physics Department, University of Colorado, Boulder, Colorado 80309.

[1] A. O. Barut and N. Ünal, J. Math. Phys. **27**, 3055 (1986).

[2] A. O. Barut, *Dynamical Groups and Generalized Symmetries in Quantum Theory* (University of Canterbury Press, Christchurch, New Zealand, 1972); A. O. Barut and R. Raçzka, *Theory of Group Representations and Applications* (World Scientific, Singapore, 1986), 2nd ed., and references therein.

[3] A. O. Barut and N. Ünal, Fortschr. Phys. **33**, 319 (1985).

0022-2488/88/010268-04$02.50