Hidden symmetry and potential group of the Maxwell fish-eye

Alejandro Frank
Instituto de Ciencias Nucleares, UNAM, México City, Mexico

François Leyvraz
Instituto de Física, UNAM, Cuernavaca, Mexico

Kurt Bernardo Wolf
Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, Universidad Nacional Autónoma de México, Cuernavaca, Mexico

(Received 26 March 1990; accepted for publication 16 May 1990)

The Maxwell fish-eye is an exceptional optical system that shares with the Kepler problem and the point rotor (mass point on a sphere) a hidden, higher rotation symmetry. The Hamiltonian is proportional to the Casimir invariant. The well-known stereographic map is extended to *canonical* transformations between of the phase spaces of the constrained rotor and the fish-eye. Their *dynamical* group is a pseudoorthogonal one that permits a succint "4π" wavization of the constrained system. The fish-eye exhibits, unavoidably, chromatic dispersion. Further, a larger conformal dynamical group contains the *potential* group, that relates the closed, inhomogeneous fish-eye system to similar, scaled ones. Asymptotically, it is related to free propagation in homogenous media.

I. INTRODUCTION

The Maxwell fish-eye is an optical medium, in principle, in any number of dimensions, whose refractive index $n(\vec{q})$ is a function $\sim(1+q^2)^{-1}$ of the distance q to the center. It is a spherically symmetric inhomogeneous system that is a *pièce de resistance* in optics textbooks and treatises[1-3] because it is very illustrative to test solution methods since the system possesses exact, closed solutions. The system was originally proposed as a problem by the Irish Academy; it asked for the refractive index of a medium that could conceivably form images in the least depth (fish eyes are notoriously flat) and Maxwell's solution was published in 1854.[4] The medium is ideal, of course, because of light injection and attenuation problems, and because of size restrictions by the physical requirement $n \geqslant 1$. Yet, this system is truly the hydrogen atom of optics, as we shall see: it possesses a manifest $SO(N-1)$ and hidden $SO(N)$ rotation symmetry groups, and $SO(N,1)$, $ISO(N)$, and $SO(N,2)$ dynamical groups.

The paths of light rays in a Maxwell fish-eye medium are closed: They are circles on planes that contain the origin, and whose points form conjugate pairs with respect to the origin. (For fixed ρ and origin, vectors $\mathbf{q}_1$ and $\mathbf{q}_2$ are conjugate when they are antiparallel and their magnitudes relate through $q_1 q_2 = \rho^2$.) In the posthumous work of Luneburg, *Mathematical Theory of Optics*,[3] a section titled "The surprising properties of an optical medium of refractive index..." shows that the circles in the Maxwell fish-eye are the *stereographic projection* of great circles on a sphere in one higher dimension. Group theory had not yet come into much vogue before 1949, when the book manuscript was assembled out of lecture notes, and contains no mention of the work of Fock[5] and Bargmann[6] on the hidden rotation symmetry of the hydrogen atom. The statement that a higher rotation symmetry is at play in the Maxwell fish-eye was made in the work by Buchdahl,[7] who mapped the constants of the fish-eye circles (plane orientation and vector to the center) onto the constants of the Kepler orbits (plane orientation and Runge–Lenz vector). These are generators of an SO(4) group under the Poisson bracket.

The Maxwell fish-eye is usually given as an example of a geometric-optics *perfect imaging instrument* because all light rays issuing from any one point in the medium will follow circle arcs that intersect at the point conjugate to the first, and the optical length of all these circle arcs between the two conjugate points is the same.[2] (Nevertheless, chromatic dispersion is not discussed in the standard texts.) Generalizations of the fish-eye, such as the Luneburg lens used in microwave antenna design[8] give the fish-eye a nontrivial practical interest.

The fish-eye is a rare instance of a "4π" optical instrument; as the hydrogen atom, it is a system worth studying for its own, group-theoretical sake. We regard it as a prime example to calibrate the Lie–Hamilton formulation of geometric and wave optics, previously used only for homogeneous optical media;[9,10] the description here includes time evolution. Section II reports a succint derivation of the Hamilton equations of motion of optics in time from the assumption only of the local validity of Snell's law. We find no extra effort for working in $N-1$ dimensions. These equations lead to the classical phase space formulation of geometrical optics, where the momentum vector is constrained to a sphere, rather than a plane as in mechanics.[11]

Section III studies a system in mechanics that is also constrained to a sphere, albeit in configuration space: the point rotor. We believe that the rotor system, rather than the Kepler problem, is the simplest mechanical $SO(N)$ model analog to the fish-eye. We write the rotation generators of the symmetry group, the $SO(N)$ Casimir invariant, and the Hamiltonian of the system as functions on phase space under Lie–Poisson brackets[12] and constrain them to their projection on the equatorial plane of the sphere. In Sec. IV we

 0022-2488/90/122757-12$03.00

introduce an extended *canonical* stereographic map of the phase spaces, the configuration part of which is the familiar sphere-on-plane 1:1 map (excluding the north pole, that maps to the point at infinity). This is an optical aberration map of the generic form of *distortion*;[13] concomitant to this, the momentum spaces of the two systems map mantaining the canonicity of the transformation. The projected rotor Hamiltonian becomes the Hamiltonian of the Maxwell fish-eye, the Casimir invariant.[14,15] We build the SO(N,1) and SO(N,2) dynamical algebra generators in Sec. V, adapting previous results for the hydrogen atom.[16] The exponentiated SO(N,2) group action is given explicitly in the geometric optics representation. This dynamical algebra includes a new *number* generator, square root of the SO(N) Hamiltonian function.

In Sec. VI we follow the same program for wave optics at a lighter pace, because spherical harmonics are well known as solutions to wave motion on a sphere.[17,18] The frequencies are discrete and given by the square roots of the Casimir operator eigenvalues. The medium can thus only sustain a discrete set of colors; sharp wave fronts will undergo chromatic dispersion. Among the concluding remarks in Sec. VII, we point out that SO(N,1) is the *potential group*,[19,20] that bridges between the fish-eye and the homogeneous medium. Finally, an SO(N,2) bundle over configuration space is suggested to describe a more general class of inhomogeneous media.

II. THE HAMILTONIAN-TIME FORMULATION OF OPTICS

We model geometric optical rays as the paths taken by points indicated by $\mathbf{q}(t)\in\Re^D$, at a time parameter t, whose velocity vector may have arbitrary direction but must be of fixed magnitude at each point of the medium,

$$\mathbf{v}(t)=\frac{d\mathbf{q}}{dt},\quad |\mathbf{v}|=\frac{c}{n(\mathbf{q})}. \tag{2.1}$$

Here, c stands for the light velocity in vacuum and $n(\mathbf{q})$ is the refractive index of the medium at the point $\mathbf{q}$ ($n=1$ characterizes vacuum). We assume that this index is a scalar function of the space coordinates only, and not of time, ray direction, or any other ray descriptor. The time needed to traverse a vanishing distance ds is $dt=n/c\,ds$. Physics asserts that c is a universal constant and that $n<1$ is unphysical.

We shall now build a vector $\mathbf{p}(t)$ tangent to the path $\mathbf{q}(t)$, i.e., parallel to $\mathbf{v}$. Snell's law is particularly transparent in suggesting the right *length* for this tangent vector $\mathbf{p}$. It is *not* $v=c/n$, as could wrongly be inferred from (2.1), but such that at any surface σ possessing a normal vector $\mathbf{\Sigma}$ and separating two otherwise homogeneous media n and n' (constant), there holds the well-known sine law:

$$n\sin\theta=n'\sin\theta', \tag{2.2}$$

with the three vectors $\mathbf{p}$, $\mathbf{p}'$, and $\mathbf{\Sigma}$ coplanar, and where as usual we denote by θ and θ' the angles between $\mathbf{\Sigma}$ and the directions $\mathbf{p}$ and $\mathbf{p}'$ of the ray before and after refraction.

Equation (2.2) may be seen as a *conservation* statement when we write each member as the magnitude of a cross product $|\mathbf{\Sigma}\times\mathbf{p}|=|\mathbf{\Sigma}||\mathbf{p}|\sin\theta$ between the (common) surface normal $\mathbf{\Sigma}$ and the vector $\mathbf{p}$ *constrained* to have length $|\mathbf{p}|=n$, or

$$\mathbf{p}\cdot\mathbf{p}=n(\mathbf{q})^2, \tag{2.3}$$

and similarly for the primed quantities. The requirement of coplanarity is the linear dependence of the three vectors: $\mathbf{p}=\alpha\mathbf{p}'+\beta\mathbf{\Sigma}$, for some real α and β. The magnitude of the cross product of $\mathbf{p}$ with $\mathbf{\Sigma}$ is consistent with (2.2) only for $\alpha=1$. A vector statement equivalent to the sine law (2.2) plus coplanarity is therefore that

$$\mathbf{p}-\mathbf{p}'=\beta\mathbf{\Sigma}. \tag{2.4}$$

Here, β is a scalar function of the vectors. If we assume $|\mathbf{\Sigma}|=1$ and decompose $\mathbf{p}=\mathbf{p}_\perp+p_\Sigma\mathbf{\Sigma}$, where $p_\Sigma=\mathbf{p}\cdot\mathbf{\Sigma}$ is the coordinate of $\mathbf{p}$ along $\mathbf{\Sigma}$, and $\mathbf{p}_\perp$ is the *conserved* component of $\mathbf{p}$ in the plane tangent to the surface. The picture we obtain of the optical medium is that for every point $\mathbf{q}\in\Re^D$ we have an $\mathcal{S}_{D-1}$ *sphere* in $\mathbf{p}$ space, that will be called the *Descartes* sphere of ray directions, whose radius depends on the point.

Indeed, Snell's law should be called Descartes' if the French philosopher, besides finding Eq. (2.2) and the requirement of coplanarity,[1] had only realized that the appropriate tangent vector $\mathbf{p}$ is *not* the velocity vector of the light corpuscule, but

$$\mathbf{v}=c/n^2\mathbf{p}, \tag{2.5}$$

so that its magnitude be consistent with (2.1) and (2.3). In the denser of two media, the ray approaches the surface normal as a particle falling in a potential step well, but actually travels slower.

We shall now derive from (2.3) and (2.4) the two Hamilton equations of motion for the light points of geometrical optics moving through inhomogeneous media, and find the function $\mathcal{H}^{\rm opt}(\mathbf{q},\mathbf{p})$ that serves as optical Hamiltonian.[11] In fact, we have done so already: Eqs. (2.1) and (2.5) compose to the first equality in

$$\frac{d\mathbf{q}}{dt}=\frac{c}{n^2}\mathbf{p}=\frac{\partial\mathcal{H}^{\rm opt}}{\partial\mathbf{p}},\quad\text{where}\quad \mathcal{H}^{\rm opt}=c\frac{\mathbf{p}\cdot\mathbf{p}}{2n^2}+\phi(\mathbf{q}), \tag{2.6}$$

where the second equality defines $\mathcal{H}^{\rm opt}$ up to an arbitrary additive function $\phi(\mathbf{q})$. The equality between the first and third terms is Hamilton's first equation. This equation, we saw, follows from the geometry of tangent vectors and the definition of $\mathbf{p}$ in (2.5).

To introduce the optical *dynamics* contained in Snell's law, we must generalize equation (2.4) for refractive indices $n(\mathbf{q})$ that possess a gradient field ∇n acting as surface normal for infinitesimal refraction $\mathbf{p}'=\mathbf{p}+d\mathbf{p}$ in a time interval dt. Equation (2.4) then becomes

$$\frac{d\mathbf{p}}{dt}=\gamma\nabla n, \tag{2.7a}$$

where we are left to determine the scalar function $\gamma(\mathbf{p},n(\mathbf{q}))$. This we do differentiating Eq. (2.3) in two different ways:

$$\frac{dn^2}{dt}=2n\nabla n\cdot\frac{d\mathbf{q}}{dt}=2n\frac{c}{n^2}\nabla n\cdot\mathbf{p} \tag{2.7b$_1$}$$

$$=2\mathbf{p}\cdot\frac{d\mathbf{p}}{dt}=2\gamma\mathbf{p}\cdot\nabla n, \tag{2.7b$_2$}$$

whence,

$$\gamma = c/n(\mathbf{q}). \tag{2.7c}$$

From (2.3) and this follows the second Hamilton vector equation of motion

$$\frac{d\mathbf{p}}{dt} = \frac{c}{n}\nabla n = -\frac{\partial \mathscr{H}^{\text{opt}}}{\partial \mathbf{q}}, \tag{2.8}$$

where the Hamiltonian function in (2.6) is forced to have $\phi(\mathbf{q}) = \text{const}$, and thus determined as

$$\mathscr{H}^{\text{opt}}(\mathbf{p},\mathbf{q}) = c[\mathbf{p}\cdot\mathbf{p}/2[n(\mathbf{q})]^2] + \text{const}, \tag{2.9}$$

is now determined up to an arbitrary additive constant. In fact, the Hamiltonian *is* constant along ray trajectories when the momentum vector $\mathbf{p}$ is everywhere on its Descartes sphere[21,22] of radius $n(\mathbf{q})$. Most important, observe that geometric *optical* Hamiltonians are constrained to have the form

$$\mathscr{H}^{\text{opt}} \sim [\text{momentum}]^2 \times [\text{scalar function of position}]. \tag{2.10}$$

The Hamilton equations of the motion are usually derived in a roundabout way from Fermat's global principle of least action, through the variational argument of the Euler–Lagrange equations. Canonical momentum is then defined as the velocity gradient of the Lagrangian and shown to participate in a condensed set of equations that are Hamilton's.[13] It is surprising that the above short derivation seems not to be known. Indeed, these arguments may be repeated *mutatis mutandis* to find the Hamiltonian evolution under translations along the optical axis [involving d/dq_i (Refs. 10,21,22)] or along the ray length [involving d/ds (Ref. 23)]. Here, we take the time dt as the infinitesimal "measuring rod." The form chosen here displays best the group-theoretical properties of the Maxwell fish-eye.

III. THE ISOTROPIC POINT ROTOR

The phase space of a nonrelativistic point mass in N dimensions is the ensemble of position coordinates $\vec{Q}\,Q = \{Q_i\}_{i=1}^N \in \Re^N$ and their conjugate momenta $\vec{P} = \{P_i\}_{i=1}^N \in \Re^N$. This is a $2N$-dimensional space where we can introduce an antisymmetric form $\{\cdot,\cdot\}$ between pairs of coordinates given by

$$\{Q_i,P_j\} = \delta_{i,j} = -\{P_j,Q_i\}, \tag{3.1}$$

$$\{Q_i,Q_j\} = 0, \quad \{P_i,P_j\} = 0. \tag{3.2}$$

This can be extended to all formal power series functions f, g, h of phase space through asking the form to be linear, to satisfy Jacobi's identity,[14] and to be a *derivation* ($\{fg,h\} = f\{g,h\} + \{f,h\}g, \{f,\text{const}\} = 0$). These properties are those of a Lie *bracket*,[12,14,15] so the coordinates of phase space serve as the basis for a Lie algebra provided we recognize the "1" in (3.1) as the central element in that algebra, with null brackets $\{1,Q_i\} = 0 = \{1,P_i\}$. This is the *Heisenberg–Weyl* algebra.[24] Classical mechanics (and geometric optics) work with the realization provided by the Poisson bracket[11]

$$\{f,g\} = \sum_{i=1}^{N}\left(\frac{\partial f}{\partial Q_i}\frac{\partial g}{\partial P_i} - \frac{\partial f}{\partial P_i}\frac{\partial g}{\partial Q_i}\right). \tag{3.3}$$

Under ordinary multiplication these functions of course commute: $fg = gf$.

The Hamilton equations for geometric optics, (2.6) and (2.8), are expressible in Poisson brackets:

$$\frac{dQ_i}{dt} = -\{\mathscr{H},Q_i\}, \quad \frac{dP_i}{dt} = -\{\mathscr{H},P_i\}. \tag{3.4a}$$

We denote $\{\mathscr{H},\cdot\}$ the Lie operator[12] associated to the observable $\mathscr{H}$. The evolution of any function $f(Q_i,P_j)$ with time t is

$$\frac{df}{dt} = \frac{\partial f}{\partial \vec{Q}}\cdot\frac{d\vec{Q}}{dt} + \frac{\partial f}{\partial \vec{P}}\cdot\frac{d\vec{P}}{dt}$$

$$= -\frac{\partial f}{\partial \vec{Q}}\cdot\frac{\partial \mathscr{H}}{\partial \vec{Q}} + \frac{\partial f}{\partial \vec{P}}\cdot\frac{d\mathscr{H}}{d\vec{P}} = -\{\mathscr{H},f\}. \tag{3.4b}$$

This is the Hamiltonian flow of phase space generated by $\mathscr{H}$.

Since $d\mathscr{H}/dt = \{\mathscr{H},\mathscr{H}\} = 0$, the trajectories in phase space $\vec{Q}(t)$, $\vec{P}(t)$, are flows along surfaces $\mathscr{H} = \text{const}$. We may use functions h other than the Hamiltonian as generators of flows: $df/ds = -\{h,f\}$, with s a length parameter along the flow lines generated by h. In particular, the flow generated by $\{Q_i,\circ\}$ is translation of phase space in the P_i direction. Similarly, the flow generated by $\{P_i,\circ\}$ is translation in the $-Q_i$ direction. Any observable f such that $\{\mathscr{H},f\} = 0$ defines surfaces $f = \text{const}$ that the flow of the Hamiltonian must respect. Finally, note that the commutator of two Lie operators $\{f,\circ\}$, $\{g,\circ\}$ is generally nonzero; in fact, it is the Lie operator of the Poisson bracket of the two functions: From the Jacobi identity we find

$$[\{f,\circ\},\{g,\circ\}] = \{f,\circ\}\{g,\circ\} - \{g,\circ\}\{f,\circ\} = \{\{f,g\},\circ\}. \tag{3.5}$$

When this quantity is zero, the Lie operators commute, and the flow generated by one function is invariant under parallel transport by the other. The Poisson bracket of the two generators is then a constant.

Linear functions of phase space close into the N-dimensional Heisenberg–Weyl algebra w_N. The independent quadratic functions are

$$A_{i,j} = P_iP_j, \tag{3.6a}$$

$$B_{i,j} = Q_iP_j + Q_jP_i, \tag{3.6b}$$

$$C_{i,j} = Q_iQ_j, \tag{3.6c}$$

$$R_{i,j} = Q_iP_j - Q_jP_i, \tag{3.7}$$

and close into the real *symplectic* algebra $\text{sp}(2N,\Re)$. The linear *plus* quadratic functions also close, the algebra is $w_N\ \text{sp}(2N,\mathscr{R})$. In particular, the $R_{i,j}$ close into the N-dimensional rotation algebra $\text{so}(N)$ that generates[14] a joint rotation of the $\vec{Q}$ and $\vec{P}$ subspaces in their i-j planes; the flow generated by $\vec{Q}\cdot\vec{P} = \frac{1}{2}\Sigma_i B_{i,j}^2$ is a radially inward flow in the $\vec{Q}$ coordinates and radially outward in the $\vec{P}$ coordinates, and leads to reciprocal scaling of the two subspaces. Flows can of course also mix the position and conjugate momentum subspaces, as those generated by P_iP_j and Q_iQ_j. Among all functions of phase space, there exist subsets that also close into Lie algebras whose vector dimension may be finite or infinite; some of them will come up for scrutiny below.

A *point rotor* is a mass point *constrained* to move on a *sphere* in configuration space

$$\vec{Q}^2 = \vec{Q}\cdot\vec{Q} = \sum_{i=1}^{N} Q_i^2 = \mathbf{Q}\cdot\mathbf{Q} + Q_N^2 = \rho^2, \tag{3.8}$$

where ρ is an arbitrary but fixed radius, and we indicate by $\mathbf{Q} = (Q_1, Q_2, \ldots, Q_{N-1})$ the first $N-1$ components of $\vec{Q} = (\mathbf{Q}, Q_N)$. On the sphere, thus,

$$Q_N^{(\sigma)} = \sigma\sqrt{\rho^2 - |\mathbf{Q}|^2}, \quad \sigma \in \{+1, 0, -1\}. \tag{3.9}$$

The *sign* σ of $Q_N^{(\pm)}$ keeps track of the two hemispheres and the $Q_N^{(0)} = 0$ equator. We shall not insist on considering separately the $\sigma = 0$ lower-dimensional manifold, but we keep in mind the natural continuity conditions between the two hemispheres. The Hamiltonian flow must leave that sphere invariant.

Those functions F of the full $\mathfrak{R}^{2N}$ phase space that have zero Poisson bracket with $\vec{Q}^2$ preserve the sphere where the point moves; among them, our purported *rotor* Hamiltonian $\mathscr{H}^{\text{rot}}$. They will generate the *symmetry* and *dynamical* group(s) of the point rotor (the latter contains the former). Functions F of phase space that have zero Poisson bracket with $\vec{Q}^2$ satisfy

$$\{F, \vec{Q}^2\} = 0, \quad \text{i.e., } 0 = \sum_{i=1}^{N} 2Q_i\{F, Q_i\} = -2\vec{Q}\cdot\frac{\partial F}{\partial \vec{P}}. \tag{3.10}$$

Among the linear and quadratic functions in (3.1), (3.6), and (3.7), only $1, Q_i, Q_iQ_j$, and $R_{i,j} = Q_iP_j - Q_jP_i$ have this property, while P_i, P_iP_j, and $B_{i,j} = Q_iP_j + Q_jP_i$ do not. This property yields a Lie algebra of functions under the Poisson bracket,[12] and its universal covering algebra (obtained by ordinary multiplication of the algebra elements) has the same property.

The *symmetry* algebra of the system will be the subset of those functions that have zero Poisson bracket also with $\mathscr{H}^{\text{rot}}$. The rotor point mass is on a sphere, with no preferred origin or direction. The set of functions $R_{i,j}$ forms a vector basis for the so(N) algebra and generates the N-dimensional rotation group SO(N). Special consideration is thus due to the SO(N) *Casimir* function of second degree in the generators (and of *fourth* degree in Q_i and P_j):

$$\Phi = \frac{1}{2}\sum_{i,j} R_{i,j}R_{i,j} = \vec{Q}^2\vec{P}^2 - (\vec{Q}\cdot\vec{P})^2. \tag{3.11}$$

In $N = 3$ dimensions, this is the squared norm of the cross product $\vec{Q}\times\vec{P}$, the angular momentum antisymmetric tensor.

Under Poisson brackets, the functions 1, Q_i, plus $C_{i,j} = Q_iQ_j$ belong to an Abelian ideal of dimension D given by 1, N, plus $\frac{1}{2}N(N+1)$ that, together with the $R_{i,j}$, form a larger "inhomogeneous" algebra $i_D$so(N) and generate a corresponding group. The Lie transformations from these functions do not affect configuration space at all: they translate *momentum* space and mix it with position. The *isotropic* point rotor needs thus one further specification: its *dynamics* must be rotation *invariant*. This means that any and only rotations of a Hamiltonian flow can be Hamiltonian flows. Hence, $\mathscr{H}^{\text{rot}}$ must be a *scalar* under rotations and so the indices in the arguments must balance. Exit the single Q_i's from consideration therefore, but retain pairs $Q_iQ_j = C_{i,j}$. Since $C_{i,j} = C_{j,i}$ but $R_{i,j} = -R_{j,i}$, a series expansion of $\mathscr{H}^{\text{rot}}$ with balanced indices can contain any number of C's, but only terms with pairs of R's.

On the sphere $|\mathbf{Q}|^2 + Q_N^2 = \rho^2$, $\Sigma_i C_{i,i} = \Sigma_i Q_iQ_i = \rho^2$ is a constant, and so are $\Sigma_{i,j}C_{i,j}C_{i,j} = \rho^4$ and $\Sigma_{i,j}C_{i,j}R_{i,j} = 0$. Higher degree polynomials of the C's and R's do not yield further independent invariants because any number of contracted factors of C, yield a single C (times a constant, since $\Sigma_j C_{i,j}C_{j,k} = \rho^2 C_{i,k}$) and any odd number of contracted R's yield a single R times a power of Φ (since $\Sigma_{j,k}R_{i,j}R_{j,k}R_{k,l} = -\Phi R_{i,l}$). The fourth-order invariant is $\Sigma R_{i,j}R_{j,k}R_{k,l}R_{l,i} = 2\Phi^2$. This reduces all higher invariants to contracted products of $\cdots RCRC\cdots$ that vanish, since for any i and m, it holds that $\Sigma_{j,k,l}R_{i,j}C_{j,k}R_{k,l}C_{l,m} = 0$. The conclusion is therefore that the point rotor Hamiltonian $\mathscr{H}^{\text{rot}}$ can be a function of Φ only, the rotation Casimir given in (3.11). Since Φ is quadratic in momentum $\vec{P}$ we may take

$$\mathscr{H}^{\text{rot}} = \omega\Phi = \omega[\vec{Q}^2\vec{P}^2 - (\vec{Q}\cdot\vec{P})^2] = E, \tag{3.12}$$

for some constant ω, E is a constant of the motion.

The constraint to the sphere $\vec{Q}^2 = \rho^2$ leaves us with a $2N$-dimensional phase space $(\vec{Q},\vec{P})$ that is too large, because there is the redundant coordinate Q_N in (3.9). Since Hamiltonian phase spaces come in even dimensions only, we should expect another constraint to be at hand. Indeed, among the quadratic functions (3.6), the rotor evolution Hamiltonian (3.12) leaves invariant the traces $\vec{Q}\cdot\vec{P} = \Sigma_i B_{i,i} = \delta$ and $\vec{P}^2 = \Sigma_i C_{i,i} = \gamma^2$ of (3.6); since there is a relation between the constants, $E = \omega[\rho^2\gamma^2 - \delta^2]$, we can choose the gauge $\delta = 0$ leaving $\gamma^2 = E/\omega\rho^2$. We denoted $\vec{Q} = (\mathbf{Q}, Q_N)$; let us similarly denote $\vec{P} = (\mathbf{P}^*, P_N)$ where $\mathbf{P}^*$ are the first $N-1$ components, so that $\vec{Q}\cdot\vec{P} = \mathbf{Q}\cdot\mathbf{P}^* + P_NQ_N = 0$ fixes

$$P_N^{(\sigma)} = -\mathbf{Q}\cdot\mathbf{P}^*/\sigma\sqrt{\rho^2 - |\mathbf{Q}|^2}. \tag{3.13}$$

What happens to the Poisson bracket structure? Disregarding σ, the position space differential under constraint is

$$d\vec{Q} = (d\mathbf{Q}, dQ_N^{(\sigma)}) = \left(d\mathbf{Q}, \frac{-\mathbf{Q}\cdot d\mathbf{Q}}{\sigma\sqrt{\rho^2 - |\mathbf{Q}|^2}}\right). \tag{3.14}$$

The *Pfaffian form*[25] (or first integral invariant of Poincaré [16(a)]) is

$$\vec{P}\cdot d\vec{Q} = \mathbf{P}^*\cdot d\mathbf{Q} + P_N\, dQ_N \tag{3.15a}$$

$$= \mathbf{P}^*\cdot d\mathbf{Q} + \frac{\mathbf{P}^*\cdot\mathbf{Q}\mathbf{Q}\cdot d\mathbf{Q}}{\rho^2 - |\mathbf{Q}|^2} \tag{3.15b}$$

$$= \mathbf{P}^*\left(\cdot 1\cdot + \frac{\cdot\,\mathbf{Q}\mathbf{Q}\,\cdot}{\rho^2 - |\mathbf{Q}|^2}\right) d\mathbf{Q} \tag{3.15c}$$

$$= \mathbf{P}\cdot d\mathbf{Q} \text{ for } \mathbf{P} = \mathbf{P}^* + \frac{\mathbf{Q}\cdot\mathbf{P}^*}{\rho^2 - |\mathbf{Q}|^2}\mathbf{Q}. \tag{3.15d}$$

The last line defines a new set of $N-1$ coordinates $\mathbf{P}$, so that the standard Pfaffian form in those $N-1$ coordinates equals the Pfaffian in the old N coordinates with the constraints. The transformation from $(\vec{Q},\vec{P})$ [restricted by (3.9) and (3.13)] to $(\mathbf{Q},\sigma,\mathbf{P})$ spaces preserves the Pfaffian and, from it, the Poisson and Lie bracket structure [(16a)],[25] *in the new variables* $(\mathbf{Q},\mathbf{P})$. [The sign σ distinguishes between two

copies of $\mathbf{Q}$ space; the $\mathbf{P}^* \mapsto \mathbf{P}$ transformation in (3.15d) is 1:1 for all points on the $|\vec{P}|$ sphere (except when $|\mathbf{Q}|^2 = \rho^2$, on the $\sigma = 0$ submanifold).] Note that it would be incorrect to simply "leave out" the Nth coordinate in the reduction from N to $(N-1)$-dimensional Poisson brackets; the projection $\vec{Q} = (\mathbf{Q}, Q_N^{(\sigma)}(\mathbf{Q})) \mapsto \mathbf{Q}$ must be accompanied by the nontrivial momentum map $\vec{P} = (\mathbf{P}^*, \vec{P}_N(\mathbf{Q},\mathbf{P}^*)) \mapsto \mathbf{P}(\mathbf{Q},\mathbf{P}^*)$. We write this map as $(\mathbf{Q},\sigma,\mathbf{P}) = (\vec{Q},\vec{P})|_{\text{rot}}$.

The Hamiltonian function (3.12) after this substitution becomes

$$\mathscr{H} = \omega \mathscr{C}, \tag{3.16a}$$

$$\mathscr{C} = \Phi|_{\text{rot}} = \rho^2 |\mathbf{P}|^2 - (\mathbf{Q}\cdot\mathbf{P})^2. \tag{3.16b}$$

Under the $|_{\text{rot}}$ map, the so(N) symmetry subalgebra generators become

$$L_{ij} = R_{ij}|_{\text{rot}} = Q_i P_j - Q_j P_i, \quad i,j = 1,\ldots,N-1, \tag{3.17}$$

$$M_i = R_{i,N}|_{\text{rot}} = -\sigma\sqrt{\rho^2 - |\mathbf{Q}|^2}\, P_i, \quad i = 1,\ldots,N-1. \tag{3.18}$$

We may verify that under the rotor map their Poisson bracket relations remain the same in the reduced $2(N-1)$-dimensional phase space $(\mathbf{Q},\mathbf{P})$, as if we had simply replaced Q_N and set $P_N = 0$.

There is a well-developed theory of constrained Hamiltonian systems of the *second class* (i.e., when the Poisson brackets of the constraints do not vanish). One may calculate the Dirac bracket [(11b,c)] between two functions of the rotor space constrained by $\xi_1 = \vec{Q}^2 - \rho^2$ and $\xi_2 = P_N$ and find the same formal result when $\xi_i = 0$. The old Poisson bracket is the replaced by the Dirac bracket, and this *equals* the Poisson bracket in the *reduced* subspace of the first $N-1$ components $\mathbf{Q}$ and $\mathbf{P}$.

In conclusion, on a homogeneous, isotropic sphere, the free motion of a point rotor in $(\vec{Q},\vec{P})$ is on sphere geodesics: arcs of great circles. The reduction to $(\mathbf{Q},\sigma,\mathbf{P})$ in effect "projects" the position coordinate of the point rotor on two copies of its equatorial plane, distinguished by the hemisphere sign σ, and with a new canonically conjugate momentum $\mathbf{P}$. In this phase space, the reduced Hamiltonian (3.16) has a natural "kinetic energy" term $|\mathbf{P}|^2$ and obeys a $(\mathbf{Q}\cdot\mathbf{P})^2$ "potential"; the trajectory jumps between the two values of the sign σ (through $\sigma = 0$) when it crosses the $|\mathbf{Q}| = \rho$ equator. We see this motion as spherical rotor motion projected on the equatorial plane.

IV. THE STEREOGRAPHIC MAP

The stereographic map is a bijection between the manifolds of the sphere $S_{N-1} \subset \Re^N$ and $\Re^{N-1} \cup \{\infty\}$ (the projection pole maps on ∞). This map was applied *deus ex machina* by Fock[5] in 1935 to the hydrogen atom Schrödinger equation in momentum representation, to obtain the hydrogenic wave functions in terms of the four-dimensional spherical harmonics. Figure 1 shows the geometry of the map for $N = 2$, between the circle S_1 and the line $\Re$, and the essentials of the general N-dimensional case. The surface of the sphere $\Sigma_{i=1}^N Q_i^2 = |\mathbf{Q}|^2 + Q_N^2 = \rho^2$ maps on the $(N-1)$-dimensional optical position space, of vectors $\mathbf{q} = \{q_i\}_{i=1}^{N-1}$. A point on the sphere that subtends the angle χ between the projecting pole and the $-Q_N$ axis, will measure an angle 2χ from the center of the sphere. The first

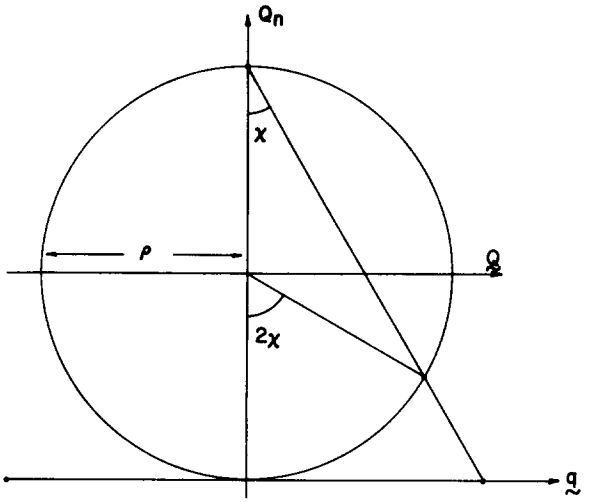

FIG. 1. The stereographic projection maps the circle on the line in $N = 2$ dimensions. The two regions $|\mathbf{Q}| < \rho$, $\sigma = \pm 1$, and their boundary $|\mathbf{Q}| = \rho$, $\sigma = 0$, map onto the full line $\mathbf{q}$; the boundary maps on $|\mathbf{q}| = 2\rho$.

appears in right triangle with sides 2ρ, $|\mathbf{q}|$, and $\sqrt{|\mathbf{q}|^2 + 4\rho^2}$, and its double in another right triangle of sides $-Q_N$, $|\mathbf{Q}|$, and ρ.

Trigonometric functions of χ and 2χ are

$$\sin\chi = \mathbf{q}/\sqrt{|\mathbf{q}|^2 + 4\rho^2}, \quad \sin 2\chi = \mathbf{Q}/\rho, \tag{4.1a}$$

$$\cos\chi = 2\rho/\sqrt{|\mathbf{q}|^2 + 4\rho^2}, \quad \cos 2\chi = -Q_N/\rho, \tag{4.1b}$$

$$\tan\chi = \mathbf{q}/2\rho, \quad \tan 2\chi = -\mathbf{Q}/Q_N. \tag{4.1c}$$

From common identities, we find

$$\mathbf{q} = \frac{2\rho\mathbf{Q}}{\rho - \sigma\sqrt{\rho^2 - |\mathbf{Q}|^2}} \tag{4.2a}$$

and

$$|\mathbf{q}|^2 = 16\rho^2 \frac{\rho + \sigma\sqrt{\rho^2 - |\mathbf{Q}|^2}}{\rho - \sigma\sqrt{\rho^2 - |\mathbf{Q}|^2}}. \tag{4.2b}$$

This is the transformation that "opens" the sphere to the plane.[10] Similarly, we find the inverse transformation

$$\mathbf{Q} = \frac{4\rho^2\mathbf{q}}{|\mathbf{q}|^2 + 4\rho^2} \tag{4.3a}$$

and

$$Q_N^{(\sigma)} = \sigma\sqrt{\rho^2 - |\mathbf{Q}|^2} = \rho\frac{|\mathbf{q}|^2 - 4\rho^2}{|\mathbf{q}|^2 + 4\rho^2}. \tag{4.3b}$$

A given point $|\mathbf{q}| < 2\rho$ $(\chi < \frac{1}{4}\pi)$ is mapped by (4.3) on $|\mathbf{Q}| < \rho$, $\sigma = -1$ (since $Q_N < 0$); the equator $|\mathbf{q}| = 2\rho$ $(\chi < \frac{1}{4}\pi)$ maps on $|\mathbf{Q}| = \rho$ and $\sigma = 0$ (with $Q_N = 0$). As $|\mathbf{q}|$ increases beyond 2ρ $(\frac{1}{4}\pi < \chi < \frac{1}{2}\pi)$, the range sweeps again through $|\mathbf{Q}| < \rho$ with $\sigma = +1$ (i.e., $Q_N > 0$). The points at zero and infinity in the plane correspond to the center of the balls $\sigma = -1$ and $\sigma = +1$ (i.e., $|\mathbf{Q}| = 0$, $Q_N = \mp\rho$), respectively. The stereographic projection is therefore a map between $\mathbf{q} \in \Re^{N-1}$ and *two open balls* $\mathbf{Q} \in \Re^{N-1}$, $|\mathbf{Q}| < \rho$, labeled by the sign σ of Q_N, whose boundaries $|\mathbf{Q}| = \rho (Q_N = 0)$ are identified.

Functions $F(\vec{Q})$ on the original sphere thus become functions $F(\vec{Q}(\mathbf{Q},\sigma))$ upon reduction of variables, and $F(\vec{Q}(\mathbf{q}))$ upon stereographic projection. In Fig. 1, where $N=2$, the $\mathbf{q}$ space is the horizontal line and $\mathbf{Q}$ space is the segment $-\rho < Q_1 < \rho$ *twice*. Figure 2 displays the $N=3$ situation showing a great circle of the S_2-sphere mapping onto a fish-eye orbit in a two-dimensional optical world.

Great circles on the $\vec{Q}$-sphere project by the stereographic map onto circles in the $\mathbf{q}$ plane.[3] Only the Nth axis is distinguished, so we may rotate the first $N-1$ coordinates such that the great circle lies in the $1-2-N$ submanifold, reducing the construction to that of Fig. 2. Further, the vector normal to the circle plane may be made to lie on the $Q_2=0$ plane, tilted by β in the Q_1 direction.

To parametrize explicitly, let us use Euler angles (β,γ) for the points of the great circle in $\vec{Q}$ space,

$$\begin{pmatrix} Q_1(\gamma) \\ Q_2(\gamma) \\ Q_N(\gamma) \end{pmatrix} = \rho \begin{pmatrix} \cos\beta\cos\gamma \\ \sin\gamma \\ -\sin\beta\cos\gamma \end{pmatrix}, \quad \begin{matrix} \beta\in[0,\pi], \\ \gamma\in S_1 = \Re \bmod 2\pi, \end{matrix} \tag{4.4}$$

where the components $3,\ldots,N-1$ are zero and omitted. In $\mathbf{Q}$ space, i.e., the equatorial plane, this draws out an ellipse of semimajor axis ρ and semiminor axis $\rho\cos\beta$. Actually, due to the twice changing sign of $Q_N^{(\sigma)}(\gamma)$, half the ellipse lies on the $\sigma=+1$ chart and half on the $\sigma=-1$ one. Through (4.2a) we find the stereographic projection of the great circle (4.4). It is

$$\begin{aligned} q_1(\gamma) &= 2\rho\frac{\cos\beta\cos\gamma}{1+\sin\beta\cos\gamma}, \\ q_2(\gamma) &= 2\rho\frac{\sin\gamma}{1+\sin\beta\cos\gamma}, \\ \gamma &\in S_1. \end{aligned} \tag{4.5}$$

Analytic geometry verifies that this is a circle of radius $2\rho\sec\beta$ with center on $-2\rho\tan\beta$. For every point $\mathbf{q}(\mathbf{Q},\sigma)\in\Re^{N-1}$ we may define its *conjugate* point[3] $\mathbf{q}^* = \mathbf{q}(-\mathbf{Q},-\sigma)$, stemming from the *antipodal* point $-\vec{Q}$ on the sphere. The vectors $\mathbf{q}$ and $\mathbf{q}^*$ are antiparallel and satisfy $|\mathbf{q}||\mathbf{q}^*| = \rho^2$.

A spherical coordinate grid with colatitude circles such as (4.5) will map onto families of *bipolar* coordinates on the plane with respect to the two poles.[17] Letting γ stand for time, it is clear that any great circle arc followed between two conjugate points will be traversed in the same time interval.[1-3] It is also clear that rotations of the sphere in the $1,\ldots,N-1$ subspace will rotate the $\mathbf{q}$ plane simultaneously; rotations into the Q_N direction produce the full family of different-sized circles that join any two fixed conjugate points in the optical $\mathbf{q}$ space. The former is the *manifest symmetry* while the latter is the *hidden symmetry* of the system.

The stereographic map and its inverse have been considered thus far as transformations of position space. They are called point transformations because the canonically conjugate momentum does not enter. In optics, such a map is called a pure distorsion. Indeed, what happens in momentum space? The coordinates $\mathbf{p}=\{p_i(\mathbf{Q},\mathbf{P})\}_{i=1}^{N-1}$ that are canonically conjugate to the components of

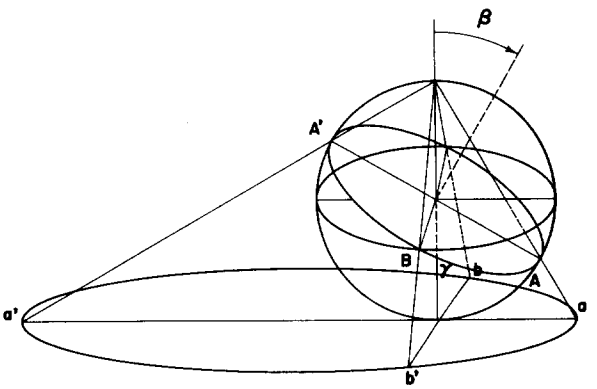

FIG. 2. The stereographic projection mapping a great circle on the $\vec{Q}$ sphere tilted by an angle β onto a circle in the $\mathbf{q}$ plane. Two pairs of antipodal-conjugate points, A-A' and B-B', are shown to map on a-a' and b-b'.

$$\mathbf{q}_i(\mathbf{Q}) = \phi(|\mathbf{Q}|)Q_i, \quad i=1,2,\ldots,N-1, \tag{4.6a}$$

may be found again from the conservation of the Pfaffian form, written $\mathbf{p}\cdot d\mathbf{q} = \mathbf{P}\cdot d\mathbf{Q}$ (Ref. 25), where now

$$\begin{aligned} d\mathbf{q}_i &= \sum_j \frac{\partial q_i}{\partial Q_j}\,dQ_j = \phi\,dQ_i + Q_i\sum_j\frac{\partial\phi}{\partial Q_j}\,dQ_j \\ &= \sum_j\left(\phi\delta_{i,j} + \phi'\frac{Q_iQ_j}{|\mathbf{Q}|}\right)dQ_j = \sum_j J_{i,j}\,dQ_j. \end{aligned} \tag{4.6b}$$

Once the Jacobian matrix $\mathbf{J}(Q)=\{J_{i,j}\}$ is known for $d\mathbf{q}=\mathbf{J}\,d\mathbf{Q}$, then $\mathbf{p}\cdot d\mathbf{q} = \mathbf{p}\cdot\mathbf{J}\,d\mathbf{Q} = \mathbf{P}\cdot\mathbf{Q}$ solves as $\mathbf{p}=\mathbf{J}^{\tau}(\mathbf{Q})^{-1}\mathbf{P}$ and $\mathbf{P}=\mathbf{J}^{\tau}(Q(q))\mathbf{p}$, where $^{\tau}$ means matrix transposition. In our case the $\{Q_iQ_j/|\mathbf{Q}|^2\}$ are idempotent matrices and we can find the canonically conjugate momentum map to be

$$\mathbf{p} = \frac{\rho-\sigma\sqrt{\rho^2-|\mathbf{Q}|^2}}{2\rho}\mathbf{P} - \frac{\mathbf{Q}\cdot\mathbf{P}}{2\rho^2}\mathbf{Q}. \tag{4.7}$$

Notice that $Q_N^{(\sigma)} = \sigma\sqrt{\rho^2-|\mathbf{Q}|^2}$ appears with its sign, so that $\mathbf{p}$ is a single-valued function over the rotor sphere. Similarly, we find the inverse transformation

$$\mathbf{P} = \frac{|\mathbf{q}|^2+4\rho^2}{4\rho^2}\left(\mathbf{p}+\frac{2\mathbf{q}\cdot\mathbf{p}}{4\rho^2-|\mathbf{q}|^2}\mathbf{q}\right). \tag{4.8}$$

We are working here only with maps between $2(N-1)$-dimensional phase spaces; the $(\vec{Q},\vec{P})$ space where the rotor motion is embedded and constrained need not be used.

Assured that the transformation between $(\mathbf{Q},\sigma,\mathbf{P})$ and $(\mathbf{q},\mathbf{p})$ spaces is canonical, we may write the $so(N)$ functions (3.17) and (3.18) in $(\mathbf{q},\mathbf{p})$. They are

$$L_{i,j} = q_ip_j - q_jp_i, \quad i,j=1,\ldots,N-1, \tag{4.9}$$

$$M_i = \rho\left[\left(1-\frac{|\mathbf{q}|^2}{4\rho^2}\right)\mathbf{p} + \frac{\mathbf{q}\cdot\mathbf{p}}{2\rho^2}\mathbf{q}\right], \quad i=1,\ldots,N-1, \tag{4.10}$$

and will close into the same algebras under the Poisson bracket in $(\mathbf{q},\mathbf{p})$ space as they did before under the Poisson bracket in $(\vec{Q},\vec{P})$ and in $(\mathbf{Q},\mathbf{P})$.

Finally, from the discussion in this section, the Casimir and Hamiltonian functions in (3.16) can be calculated replacing $\mathbf{q}(\mathbf{Q},\mathbf{P})$ and $\mathbf{p}(\mathbf{Q},\mathbf{P})$. They are

$$\mathscr{C} = \left(\sum_{i<j=2}^{N-1} L_{ij}^2 + \sum_{i=1}^{N-1} M_i^2\right)$$

$$= \rho^2\left(1+\frac{|q|^2}{4\rho^2}\right)^2 |\mathbf{p}|^2 = \frac{1}{\omega}\mathscr{H}(\mathbf{q},\mathbf{p}) = \frac{\mathbf{E}}{\omega}. \quad (4.11)$$

Now we compare this "stereographically projected" rotor Hamiltonian $\mathscr{H}$ with the generic *optical* Hamiltonian $\mathscr{H}^{\text{opt}} = c|\mathbf{p}|^2/n(\mathbf{q})|^2$ in Eq. (2.9). Their dependence on the squared momentum $|\mathbf{p}|^2$ is the same and they coincide only when the refractive index of the optical medium is

$$n(\mathbf{q}) = \frac{n_0}{1+|\mathbf{q}|^2/4\rho^2}, \quad n_0 = n(\mathbf{0}) = \frac{1}{\rho}\sqrt{\frac{c}{2\omega}}. \quad (4.12)$$

This is the refractive index that characterizes the Maxwell fish-eye.

V. THE FISH-EYE so(*N*,2) DYNAMICAL ALGEBRA

The realization of the so(N) algebra in (4.9) and (4.10) is well known from the theory of the hydrogen atom.[5,6,15,16] The momentum space of that system is the stereographic projection of the Fock sphere, where rotor momentum moves. Here, it is configuration space that maps under the stereographic projection. The earliest reference to (the Fourier transform [$\mathbf{q}\mapsto\mathbf{p},\mathbf{p}\mapsto-\mathbf{q}$] of) the so($N$,2) algebra written here seems to be Ref. 16. It is perhaps best known from the book by Wybourne,[26] who quotes the result of Barut and Bornzin.[27] Here, we shall use these results to examine the exponentiation to the SO(N,2) *group* of transformations of the Maxwell fish-eye optical phase space. This phase space is, we recall, $\mathbf{q}\in\Re^{N-1}$ and $\mathbf{p}$ constrained, for each point in $\mathbf{q}$ space, to lie on the Descartes ray-direction sphere (2.3), i.e., a sphere $S_{N-2}\subset\Re^{N-1}$ of radius

$$p(\mathbf{q}) = |\mathbf{p}| = \sqrt{\mathscr{C}}/(\rho[1+|\mathbf{q}|^2/4\rho^2]),$$

as demanded by the constancy of (4.11).

The set of functions L_{ij} in (4,7), for $i,j = 1,2,\ldots,N-1$, that generate ordinary joint rotations of $\mathbf{q}$ and $\mathbf{p}$ spaces, close into an so($N-1$) algebra that integrates to an SO($N-1$) group. These transformations map fish-eye orbits onto similar fish-eye orbits, rotated around the origin. This is the *manifest symmetry* group of the Maxwell fish-eye.

The N functions M_i, $i = 1,2,\ldots,N-1$ transform as a vector under SO($N-1$) and, together with the L_{ij}'s, are generators of an so(N) algebra. The corresponding group is SO(N); it maps any given fish-eye orbit onto all other possible orbits in the same medium. Fish-eye orbits can thus be made to change their radius and center. This is the hidden symmetry group of the fish-eye. Still a symmetry, though.

Enter dynamics. We may calculate that

$$\{M_i,q_j\} = -\delta_{ij}\rho(1-|\mathbf{q}|^2/4\rho^2) - q_iq_j/2\rho, \quad (5.1a)$$

$$\{M_i,p_j\} = L_{ij}/2\rho + \delta_{ij}\mathbf{q}\cdot\mathbf{p}/2\rho. \quad (5.1b)$$

The first expression shows that under the integrated action of $\{\mathbf{a}\cdot\mathbf{M},\circ\}$, $\mathbf{q}$ will map (nonlinearly) into $\mathbf{q}'(\mathbf{q},\mathbf{a})$. The second expression shows that if we consider p_j, $\mathbf{q}\cdot\mathbf{p}$, and

$$\{M_i,\mathbf{q}\cdot\mathbf{p}\} = \rho\left[-\left(1+\frac{|\mathbf{q}|^2}{4\rho^2}\right)p_i + \frac{\mathbf{q}\cdot\mathbf{p}}{2\rho^2}q_i\right], \quad (5.2)$$

then further Poisson brackets of these functions will close into an algebra larger than so(N). To identify this algebra we recall that the Cartesian basis of so(N,M) generators satisfy

$$\{\Lambda_{ij},\Lambda_{k,l}\} = g_{j,k}\Lambda_{l,i} + g_{i,l}\Lambda_{k,j} + g_{j,l}\Lambda_{i,k} + g_{i,k}\Lambda_{j,l}, \quad (5.3a)$$

where

$$g_{j,k} = \begin{cases} 1, & j=k\leqslant N, \\ -1, & N+1\leqslant j=k\leqslant N+M, \\ 0, & \text{otherwise.} \end{cases} \quad (5.3b)$$

We take $\Lambda_{ij} = -\Lambda_{j,i}$ for i,j both in the range $(1,\ldots,N)$ or both in $(N+1,\ldots,N+M)$, and $+\Lambda_{j,i}$ otherwise.

We write the so(N,1) generators in the following way:

$$L_{ij} = \Lambda_{ij}, \quad (5.4a)$$

$$M_i = \Lambda_{i,N}, \quad i,j = 1,2,\ldots,N-1, \quad (5.4b)$$

$$K_i = \Lambda_{i,N+1} = M_i - 2\rho p_i$$

$$= \rho\left[-\left(1+\frac{|\mathbf{q}|^2}{4\rho^2}\right)p_i + \frac{\mathbf{q}\cdot\mathbf{p}}{2\rho^2}q_i\right], \quad (5.4c)$$

$$K_N = \Lambda_{N,N+1} = -\mathbf{q}\cdot\mathbf{p}. \quad (5.4d)$$

We recognize the "noncompact" generators to be the K's, formally because of the minus sign in $\{K_i,K_j\} = -L_{ij}$, and manifestly because $-\mathbf{q}\cdot\mathbf{p}$ generates unbounded magnifications of configuration space. On smooth functions f of phase space,

$$\exp\beta\{K_N,\circ\}:f(\mathbf{q},\mathbf{p})\mapsto f(e^{-\beta}\mathbf{q},e^{\beta}\mathbf{p}), \quad \beta\in\Re. \quad (5.5)$$

We note that this action is no longer an invariance transformation of the fish-eye Hamiltonian (4.9) but, for $\mathscr{H}^{\text{fish-eye}}$ as a function of rotor radius ρ,

$$\exp-\beta\{K_N,\circ\}:\mathscr{H}(\rho)^{\text{fish-eye}}$$

$$\mapsto\omega\rho^2(1+e^{2\beta}|\mathbf{q}|^2/4\rho^2)^2e^{-2\beta}|\mathbf{p}|^2 = \mathscr{H}(e^{-\beta}\rho)^{\text{fish-eye}}. \quad (5.6)$$

In particular, we note that the radius ρ of the rotor sphere dilates to infinity for $\beta\to-\infty$. If we set $\omega = \frac{1}{2}cn_o^{-2}\rho^{-2}$, then by (4.10) we map the Maxwell fish-eye Hamiltonian $\mathscr{H}(\rho)^{\text{fish-eye}}$ onto $\mathscr{H}(\infty)^{\text{fish-eye}} = n_o$, the optical Hamiltonian of an infinite homogeneous medium.

The other K_i, $i = 1,2,\ldots,N-1$ will produce SO(N)-rotated versions of this action. Of these we wish to remark a direct physical interpretation. The SO(N,1) generators in (5.4c) may *also* be obtained through the standard *deformation* formula[14,28] as $K_i = \{\mathscr{C},p_i\}$, in general with a summand τp_i, $\tau = \text{const}$. This suggests $\omega^{-1}\{\mathscr{H}^{\text{fish-eye}},p_i\} = \omega^{-1}\,dp_i/dt$, to be a ray "acceleration" vector.

We observe that the functions p_i are linear combinations of so(N,1) generators, viz., $p_i = (M_i - K_i)/2\rho$. Thus a second visible noninvariance transformation of the optical fish-eye Hamiltonian is

$$\exp\Sigma_i a_i\{M_i - K_i,\circ\}:\mathscr{H}(\mathbf{q},\mathbf{p})^{\text{fish-eye}}\mapsto\mathscr{H}(\mathbf{q}-2\rho\mathbf{a},\mathbf{p})^{\text{fish-eye}}, \quad (5.7)$$

i.e., the map is to *another* fish-eye whose origin is at $\mathbf{a}$ instead of the origin. The algebra generated by L_{ij}'s and p_i's is the

Euclidean algebra $\mathrm{iso}(N-1)\subset\mathrm{so}(N,1)$.

The group $SO(N,1)$ thus obtained therefore contains not only the symmetry group of the Maxwell fish-eye, but also the transformations between all possible such fish-eyes, translated and dilated, up to and including asymptotically the homogeneous medium. This is the *potential group* of the Maxwell fish-eye. Potential *algebras* of the family $\mathrm{so}(M,N)$ were used in Refs. 19 and 20, to relate the quantum Pöschl–Teller and other mostly one-dimensional potentials to the free particle. That this approach also serves optical systems, is in principle remarkable.

A search for further functions closing under Poisson brackets with the generators of $\mathrm{so}(N,1)$ is rewarded when we introduce the function $p=\sqrt{\mathbf{p}\cdot\mathbf{p}}=|\mathbf{p}|$. We may then verify that an algebra is formed by the previous $\mathrm{so}(N,1)$ generators, plus

$$H_i=\Lambda_{i,N+2}=q_ip,\quad i=1,2,\ldots,N-1, \tag{5.8a}$$

$$H_N=\Lambda_{N,N+2}=H_{N+1}-2\rho p=-\rho(1-|\mathbf{q}|^2/4\rho^2)p, \tag{5.8b}$$

$$\mathcal{N}=H_{N+1}=\Lambda_{N+1,N+2}=\rho(1+|\mathbf{q}|^2/4\rho^2)p=+\sqrt{\mathcal{C}}. \tag{5.8c}$$

We note prominently that $\mathcal{N}$ in (5.8c) is the square root of the $\mathrm{so}(N)$ Casimir function $\mathcal{C}$ in (4.9b). It is a compact generator of an $\mathrm{so}(N,2)$ algebra. In the hydrogen-atom system,[16] this 'last' generator is the *number* operator.

Let us examine finally the integrated group action generated by $\mathcal{N}=H_{N+1}=\Lambda_{n+1,N+2}$. This is a *sui generis* evolution of the fish-eye system. From (5.3),

$$\exp s\{\mathcal{N},\circ\}:\Lambda_{ij}=\begin{cases}\Lambda_{ij}, & 1\leqslant i,j\leqslant N,\\ K_i\cos s-H_i\sin s, & 1\leqslant i\leqslant N, j=N+1,\\ K_i\sin s+H_i\cos s, & 1\leqslant i\leqslant N, j=N+2.\end{cases} \tag{5.9}$$

Hence, for $p_i=(M_i-K_i)/2\rho$ and $p=(H_{N+1}-H_N)/2\rho$,

$$\exp s\{\mathcal{N},\circ\}:p_i=\left[1-\frac{1-\cos s}{2}\left(1+\frac{|\mathbf{q}|^2}{4\rho^2}\right)\right]p_i+\left[p\sin s+\frac{1-\cos s}{4\rho^2}\mathbf{q}\cdot\mathbf{p}\right]q_i, \tag{5.10a}$$

$$\exp s\{\mathcal{N},\circ\}:p=\left[\cos s+\frac{1-\cos s}{2}\left(1+\frac{|\mathbf{q}|^2}{4\rho^2}\right)\right]p-\frac{1}{2\rho}\mathbf{q}\cdot\mathbf{p}\sin s. \tag{5.10b}$$

Finally, for $q_i=H_i/p$,

$$\exp s\{\mathcal{N},\circ\}:q_i=\frac{[p\cos s+\mathbf{q}\cdot\mathbf{p}/2\rho\sin s]q_i-\rho(1+|\mathbf{q}|^2/4\rho^2)\sin s\;\;p_i}{\left[\cos s+\frac{1}{2}(1-\cos s)(1+|\mathbf{q}|^2/4\rho^2)\right]p-\frac{1}{2}\mathbf{q}\cdot\mathbf{p}/\rho\sin s}. \tag{5.11}$$

The evolution parameter s is along the flow lines $\nu=\rho(1+|\mathbf{q}|^2/4\rho^2)p=\text{const}$, or $H^{\text{fish-eye}}=\text{constant}$. Thus we find the optical evolution generated by the Hamiltonian $H^{\text{fish-eye}}=\omega\mathcal{N}^2$ through observing that

$$\exp t\{H^{\text{fish-eye}},\circ\}=\exp(\omega t/2\nu)\{\mathcal{N},\circ\}. \tag{5.12a}$$

The result is then given by (5.10) and (5.11), replacing the parameter s by the time t through

$$s=\omega t/2\rho(1+|\mathbf{q}|^2/4\rho^2)p. \tag{5.12b}$$

This is the transformation of phase space along the orbits of the fish-eye system.

VI. WAVIZATION OF THE MAXWELL FISH-EYE

In this section we shall use another well-known realization of the symmetry and potential algebras and groups that describes wave optics. We will "wavize" (or "ondulate"?) the Maxwell fish-eye by a method analogous to the dynamical quantization of mechanical systems.[29]

The scalar wave equation for the field amplitude $\Phi(\vec{Q},t)$ in a homogeneous N-dimensional medium $\vec{Q}\in\mathfrak{R}^N$, of refractive index n_0, is

$$\sum_{i=1}^{N}\frac{\partial^2\Phi(\vec{Q},t)}{\partial Q_i^2}=\left(\frac{n_0}{c}\right)^2\frac{\partial^2\Phi(\vec{Q},t)}{\partial t^2}. \tag{6.1}$$

This equation may be put in evolution form (i.e., with first-order time derivative) through doubling the function space,

$$\begin{pmatrix}0 & 1\\ (c/n_0)^2\Delta & 0\end{pmatrix}\begin{pmatrix}\Phi\\ \dot{\Phi}\end{pmatrix}=\frac{\partial}{\partial t}\begin{pmatrix}\Phi\\ \dot{\Phi}\end{pmatrix}, \tag{6.2a}$$

where the space Laplacian Δ is

$$\Delta=\sum_{i=1}^{N}\frac{\partial^2}{\partial Q_i^2}. \tag{6.2b}$$

The first component equation in (6.2a) defines $\dot{\Phi}(\vec{Q},t)=\partial\Phi(\vec{Q},t)/\partial t$, and the second then reproduces (6.1). The solutions of the wave equation can be expressed in terms of the initial conditions $(\Phi_0,\dot{\Phi}_0)$ at $t=0$, through integration of the one-parameter evolution group. This is

$$\begin{pmatrix}\Phi(\vec{Q},t)\\ \dot{\Phi}(\vec{Q},t)\end{pmatrix}=\exp t\begin{pmatrix}0 & 1\\ (c/n_0)^2\Delta & 0\end{pmatrix}\begin{pmatrix}\Phi_0(\vec{Q})\\ \dot{\Phi}_0(\vec{Q})\end{pmatrix}, \tag{6.3}$$

provided the refractive index n_0 is independent of time—a good general assumption.

When light is of a definite *color*, i.e., when the time behavior of the wave function is that of a single Fourier component ν,

$$\Phi^{(\nu)}(\vec{Q},t)=\Phi_0(\vec{Q})e^{i\nu t}, \tag{6.4}$$

then the time-independent wave equation for a homogeneous medium factorizes as

$$\Delta\Phi_0(\vec{Q}) = \left(\frac{1}{|\vec{Q}|^{N-1}}\frac{\partial}{\partial|\vec{Q}|}|\vec{Q}|^{N-1}\frac{\partial}{\partial|\vec{Q}|}+\frac{\tilde{\mathscr{C}}}{|\vec{Q}|^2}\right)\Phi_0(\vec{Q}) = -\left(\frac{\nu n_0}{c}\right)^2\Phi_0(\vec{Q}); \tag{6.5}$$

here we have further factorized into radial and angular variables

$$\tilde{\mathscr{C}} = -\frac{1}{2}\sum_{j,k}^{N}\tilde{\Lambda}_{j,k}^2 \tag{6.6a}$$

$$\tilde{\Lambda}_{j,k} = i\left(Q_j\frac{\partial}{\partial Q_k}-Q_k\frac{\partial}{\partial Q_j}\right). \tag{6.6b}$$

Notice that in (6.5), the factor $(\nu n_0/c)^2 \geqslant 0$ appears where the familiar energy eigenvalue $2mE/\hbar^2$ appears in the time-independent Schrödinger equation of quantum mechanics.

The wave equation on the sphere $S_{N-1} \subset \Re^N$ is obtained from (6.1) after separation in spherical coordinates. This reduces the equation to

$$\tilde{\mathscr{C}}\Phi(\vec{Q}) = \left(\frac{n_0\nu\rho}{c}\right)^2\Phi(\vec{Q}). \tag{6.7}$$

The operators (6.6b) are well known $\mathscr{L}_2(\Re^N)$ self-adjoint realizations of the rotation algebra and group generators. Only $N-1$ variables among the Q_i are independent because, as in (3.9), $Q_N = \sigma\sqrt{\rho^2-|\mathbf{Q}|^2}$. Through the stereographic projection (4.2) and (4.3) we may now map the Q_i coordinates of the ρ sphere in (6.6) onto the $\mathbf{q}$ plane $\Re^{N-1}$ where the Maxwell fish-eye lives. The chain rule for (4.2a) would yield

$$\frac{\partial}{\partial Q_i} = \left(1+\frac{|\mathbf{q}|^2}{4\rho^2}\right)\left(\frac{\partial}{\partial q_i}+\frac{2q_i}{4\rho^2-|\mathbf{q}|^2}\mathbf{q}\cdot\frac{\partial}{\partial\mathbf{q}}\right), \tag{6.8}$$

for $i = 1,2,\ldots,N-1$. Compare with the geometrical (classical) expression (4.8) for the canonical conjugate P_j: functions on the ρ sphere no longer have an independent coordinate Q_N, and $\partial/\partial Q_N$ acts as zero and plays no further role.

While in $\Re^N$, the independent formal operators, Q_i and $\partial/\partial Q_j$ close into the N-dimensional Heisenberg–Weyl algebra, after restriction to S_{N-1} and subsequent stereographic projection, they do so only for $i,j = 1,2,\ldots,N-1$. The operators (6.8) generate translations that do not leave the sphere S_{N-1} invariant, so they cannot be exponentiated alone; they are not self-adjoint on the space of functions on the sphere. However, the rotation generators $\tilde{\Lambda}_{i,j}$ in (6.6b) *are* self-adjoint, and hence valid operators on the sphere. Other valid operators are the Q_i, $i = 1,2,\ldots,N-1$, and products or uniformly convergent series thereof.

The Casimir operator $\tilde{\mathscr{C}}$ appears in (6.7); its eigenvalues on the space of single valued and bounded functions are

$$\lambda = l_N(l_N+N-2), \quad l_N = 0,1,2,\ldots. \tag{6.9a}$$

The light colors ν that the compact space S_{N-1} can sustain are thus limited to the discrete frequencies

$$\nu_N = \frac{c}{n_0\rho}\sqrt{l_N(l_N+N-2)}, \quad l_N = 0,1,2,\ldots. \tag{6.9b}$$

We know[18] the Hilbert space $\mathscr{L}^2(S_{N-1})$ of Lebesgue square-integrable functions $\Phi(\vec{Q})$ over the sphere S_{N-1}. These functions are first mapped on the two functions $\overline{\Phi}_{\sigma=\pm1}(\mathbf{Q}) = \Phi(\vec{Q})$ on the balls in $\Re^{N-1}$ where $|\mathbf{Q}|<\rho$, as we saw in Sec. IV. Then we proceed through the stereographic projection on wave functions $\phi(\mathbf{q}) = \overline{\Phi}_\sigma(\mathbf{Q})$, with $\mathbf{q}(\mathbf{Q},\sigma)\in\Re^{N-1}$ as given in (4.2). The $\mathscr{L}^2(S_{N-1})$ inner product of two functions Φ, Ψ is thus

$$(\Phi,\Psi)_{\mathscr{L}^2(S_{N-1})} = \int_{S_{N-1}} d^{N-1}\Omega(\vec{Q})\Phi(\vec{Q})^*\Psi(\vec{Q}) \tag{6.10a}$$

$$= \sum_{\sigma=\pm1}\frac{1}{\rho^{N-2}}\int_{|\mathbf{Q}|<\rho}\frac{d^{N-1}\mathbf{Q}}{\sigma\sqrt{\rho^2-|\mathbf{Q}|^2}}\overline{\Phi}_\sigma(\mathbf{Q})^*\overline{\Psi}_\sigma(\mathbf{Q}) \tag{6.10b}$$

$$= \frac{1}{\rho^{N-1}}\int_{\Re^{N-1}}\frac{d^{N-1}\mathbf{q}}{(1+|\mathbf{q}|^2/4\rho^2)^{N-1}} \times\Phi(\vec{Q}(\mathbf{q}))^*\overline{\Phi}(\vec{Q}(\mathbf{q})) \tag{6.10c}$$

$$= (\phi,\psi)_{\text{fish-eye}} = \int_{\Re^{N-1}} d^{N-1}\mathbf{q}\phi(\mathbf{q})^*\psi(\mathbf{q}), \tag{6.10d}$$

where

$$\phi(\mathbf{q}) = \frac{\Phi(\vec{Q}(\mathbf{q}))}{\sqrt{(1+|\mathbf{q}|^2/4\rho^2)^{N-1}}}, \tag{6.11}$$

and Ψ and ψ are bound by a similar relation.

Under this inner product, symmetry transformations are unitary and their infinitesimal generators (6.6b) are self-adjoint. When we use the customary inner product form $(\phi,\psi)_{\text{fish-eye}}$ in (6.10d) for a "flat" space of measure $d^{N-1}\mathbf{q}$, the so(N) generators $\hat{L}_{i,j}$ and $\hat{M}_i$ are the Schrödinger quantization of the "classical functions" in (3.17) and (3.18) and (4.9) and (4.10) [i.e., through the replacements $q_i \mapsto q_i\cdot$, (multiplication by q_i), and $p_i \mapsto \hat{p}_i = -i\partial/\partial q_i$]. Because the functions involved are linear in the components of p_i, there is no operator-ordering ambiguity in this case: any quantization rule that guarantees self-adjointness under the inner product $\int_{\Re^{N-1}} d^{N-1}\mathbf{q}\cdots$ Ref. 24 yields

$$f(q_i)p_j \mapsto \tfrac{1}{2}\{f(\hat{q}_i),\hat{p}_j\}_+ = \tfrac{1}{2}[f(\hat{q}_i)\hat{p}_j+\hat{p}_jf(\hat{q}_i)] = -if(q_i)\frac{\partial}{\partial q_j}+\frac{1}{2}i\frac{\partial f(q_i)}{\partial q_j}.$$

This allows us to write the Maxwell fish-eye *dynamical* generators $\hat{K}_i$ in (5.4), independent of any ordering rule.

The optical fish-eye Hamiltonian is the Casimir operator of SO(N) built in (3.17) and (4.9). By itself, as a function in ($\mathbf{Q}$,$\mathbf{P}$) or ($\mathbf{q}$,$\mathbf{p}$), the Hamiltonian function would be subject to ordering-rule ambiguities;[24] however, as Casimir operator, $\tilde{\mathscr{C}}$ is the sum of squares of the operators (6.6a); defined thus, the Hamiltonian is unique and independent of the quantization scheme. All higher-order Casimir invariants are zero: the sphere S_{N-1} can only support the *totally symmetric* representations of SO(N). The number of independent Maxwell fish-eye states that are degenerate for some l_N is given by the branching rules of the so(N) representations. The representation *row* indices are provided by the canonical basis,[14] and given as a $(N-1)$-plet $\{l_N,l_{N-1},\ldots,l_2\}$, with the integer labels l_j bound by

$l_N \geqslant l_{N-1} \geqslant \cdots \geqslant l_3 \geqslant |l_2|$. The count is $2l_2 + 1$ for the spherical harmonics on the surface of the ordinary sphere S_2, and $(l_3 + 1)^2$ for the curved S_3 space that we use in the model of the "physical" Maxwell fish-eye.

In the plane two-dimensional optical world projected in Fig. 2 ($N = 3$), the description of the wave patterns in the Maxwell fish-eye is easy: The labels are the usual $\{l,m\}$, and the wave solutions are in the linear span of the ordinary solid spherical harmonic basis $\{\mathscr{Y}_{l,m}(\vec{Q})\}_{m=-l}^{l}$ (Ref. 18), each vibrating with an angular frequency (6.9b), namely

$$\nu(l) = (c/n_0\rho)\sqrt{l(l+1)}, \quad l = 0,1,2,\ldots . \tag{6.12a}$$

The time evolution of a linear combination of harmonics with coefficients $f_{l,m}$ is then

$$\Phi(\vec{Q},t) = \sum_{l=0}^{\infty} e^{i\nu(l)t} \sum_{m=-l}^{l} f_{l,m}\mathscr{Y}_{l,m}(\vec{Q}). \tag{6.12b}$$

These functions can be visualized as a pattern of light with intensity $(\Phi,\Phi)_{\mathscr{L}^2(S_2)}$ on the ordinary S_2 sphere of radius ρ. Light of a given color number l has only $2l + 1$ distinct wave patterns labeled by $m = 0, \pm 1,\ldots, \pm l$. Rotations of the sphere will mix m's, maintaining the linear subspaces l invariant. When the stereographic projection (4.3) is applied, with the weight (obliquity) factor given in (6.11), the projections of the spherical harmonics on the optical $\mathbf{q}$ space will provide an $\mathscr{L}^2(\Re^2)$-orthogonal basis for the Maxwell fish-eye solutions. These are

$$\Upsilon_{l,m}(\mathbf{q},t) = \frac{Y_{l,m}(\mathbf{q}/(1 + |\mathbf{q}|^2/4\rho^2))}{1 + |\mathbf{q}|^2/4\rho^2} \times \exp(ict\sqrt{l(l+1)}/n_0\rho), \tag{6.13}$$

where we have written the solid spherical harmonic constrained to the sphere in Cartesian coordinates as $\mathscr{Y}_{l,m}(\vec{Q})|_{S_2} = \mathscr{Y}_{l,m}(\mathbf{Q},\sigma\sqrt{\rho^2 - |\mathbf{Q}|^2}) = \rho^l Y_{l,m}(\mathbf{Q},\sigma)$.

Consider first the "extreme" $m = \pm l$ wave patterns $\Upsilon_{l,\pm l}(\mathbf{q}(\beta,\gamma),t)$, using polar angles (β,γ) on the sphere. The functions behave as $\sin^l\beta \exp i(\pm l\gamma + ct/n_0\rho\sqrt{l(l+1)})$ and will exhibit l moving nodal meridians with a braid of maxima at the equator $\beta = \pi/2$. The pattern is a function of $\gamma + \omega_l t$, where ω_l is the angular velocity of the sphere, $\omega_l = (c/n_0\rho)\sqrt{1 + 1/l}$, $l = 1,2,\ldots$. That rotating light pattern will project on the optical plane as the circular motion of waves in the fish-eye, with l nodes as spokes in a rigid rotating wheel. The belt of light maxima may also rotate on an inclined axis; it will then project its equatorial braid on an off-center circle on the fish-eye plane, the nodal meridians will project on circular nodes that cross through the two conjugate points that are images of the new rotation poles. These 'circle-of-light' rotating solutions, we surmise, are the best wave analogs of the geometric light orbits, such as that of Fig. 2.

We note that, inevitably, *chromatic dispersion* takes place: $\omega_l \sim \sqrt{1 + 1/l}$ is not independent of l, as it is in a homogeneous optical medium where wave velocity is independent of wave number. For $l = 0$, $\Upsilon_{0,0}(\mathbf{q}) = \text{const}/(1 + |\mathbf{q}|^2/4\rho^2)$ in (6.13). For growing l, $|\omega_l|$ decreases monotonically from $\omega_1 = c\sqrt{2}/n_0\rho$ down to $\omega_\infty = c/n_0\rho$. Asymptotically with l, we see that the surface of the sphere of radius ρ moves at the equator with velocity $v = \omega_\infty = c/n_0$.

Another set of harmonics that are easy to visualize, are the $m = 0$ solutions $\mathscr{Y}_{l,0}(Q_3)$. They contain a Legendre polynomial $P_l(\cos\beta)$, that has l nodal circles (only one can be the equator great circle); they are independent of the longitude angle γ. These linear multipole standing-wave solutions have their global absolute maxima at the two sphere poles, with a relative sign $(-1)^l$ between the two, and their single, *sui generis* vibration frequency $\nu(l)$ in (6.12a). When this vibration mode is tilted by SO(3) to any angle and projected on $\mathbf{q}$ space, the strongly elongating polar regions will map on a conjugate pair of pulsating light zones in the Maxwell fish-eye. They will be in or out of phase according to the parity of l. A Dirac δ flash at some point of the fish-eye, or at a closed wave front line $\mathbf{q}(\beta_0,\gamma)$, $\gamma \in S_1$, will decompose by

$$\delta(\cos\beta - \cos\beta_0) \sim \sum_{l=0}^{\infty} P_l(\cos\beta)\,P_l(\cos\beta_0)$$

into a weakly convergent series of the above conjugate-pair "standing waves."

There is dispersion again. Under time evolution, the coefficients $P_l(\cos\beta_0)$ in the series will be multiplied by $e^{i\nu(l)t}$, $\nu(l) \sim \sqrt{l(l+1)}$, whose periods are incommensurable. Thus although the optical path between two conjugate points is equal along any circle arc joining the points, and wave fronts are well defined,[1] we see that the Maxwell fish-eye is *not quite* a perfect imaging device[2] in the sense that it cannot forestall the chromatic dispersion that will smear out any pulse. This is in contrast with optics in a homogeneous space, where spherical δ wave fronts propagate as such in odd dimensions (and develop a trailing wake in even dimensions[17]). Although we can work with Dirac δ's on the S_2 sphere, they are not *eigen*functions of any rotation generators.

In the plane optical world of Fig. 2, the manifest symmetry generator is $\hat{L}_{1,2}$ and the hidden symmetry generators are $\hat{M}_1$ and $\hat{M}_2$. The extra generators of the dynamical algebra so(3,1) are $\hat{K}_1$, $\hat{K}_2$, and $\hat{K}_3$, given by (5.4c) and (5.4d). This enlarged linear space of operators may be used to define other bases for the polychromatic, wavized fish-eye. We refer to Eqs. (5.4c) to choose the two commuting operators

$$\hat{p}_i = \frac{1}{2\rho}(\hat{M}_i - \hat{K}_i) = -i\frac{\partial}{\partial q_i}, \quad i = 1,2, \tag{6.14}$$

that are generators of the Euclidean algebra iso(3) together with the symmetry generators. These define the plane-wave generalized basis of the $\mathscr{L}^2(\Re^2)$ with the inner product $(\cdot,\cdot)_{\text{fish-eye}}$ in (6.10d). Such solutions quickly loose their shape under Maxwell fish-eye evolution because they are eigenfunctions of operators that do not commute with the driving Hamiltonian. Their time evolution may be calculated group theoretically through the transformation (5.10a) or as overlap coefficients between the elliptic (l,m) and parabolic (p_1,p_2) subgroups of the SO(3,1) group that we have realized on the $\Re^2$ plane.

Extending further the dynamical algebra group to the conformal SO(3,2) by the geometric-optics generators $H_i = q_i p$, $i = 1,2$, $H_3 = \mathscr{N} - 2\rho p$, and $H_4 = \mathscr{N}$

$=\rho(1+|q|^2/4\rho^2)p$ [generically SO(N,2) for $(N-1)$-dimensional fish-eyes], we have integral operators: the scalar root of the Laplacian, $p=\sqrt{p_1^2+p_2^2}$. Its action $\hat{p}$ on smooth, properly decreasing functions $\phi(\mathbf{q})$ is through a formally divergent kernel (as a Dirac δ' derivative) given by

$$(\hat{p}\phi)(\mathbf{q})=\int_{\Re^2} d^2\mathbf{q}'\Pi(\mathbf{q}-\mathbf{q}')\phi(\mathbf{q}'),$$

$$\Pi(\mathbf{q})=\frac{i}{4\pi^2}\int_{\Re^2} d^2\mathbf{p}|\mathbf{p}|\exp i\mathbf{p}\cdot\mathbf{q}. \tag{6.15}$$

This scalar root operator may enter commutators and symmetrized products. Indeed, the Heisenberg–Weyl commutators of

$$\hat{H}_i=\tfrac{1}{2}(\hat{q}_i\hat{p}+\hat{p}\hat{q}_i) \text{ and } \hat{\mathcal{N}}-\rho\hat{p}=\tfrac{1}{2}(|\hat{q}|^2\hat{p}+\hat{p}|\hat{q}|^2)$$

close, with the rest of the SO(3,1) generators, into the Lie algebra of SO(3,2) in Eqs. (5.3). In this realization we have the SO(3) Casimir operator written as $\hat{\mathcal{N}}(\hat{\mathcal{N}}+1)$. Generally, for SO($N$,2), it is $\hat{\mathcal{N}}(\hat{\mathcal{N}}+N-2)$.

VII. CONCLUDING REMARKS

When a "physical" three-dimensional optical medium is homogeneous, the symmetry of the system is the Euclidean algebra iso(3).[9] In this case there are no additional, hidden symmetries. The optical Hamiltonian is then also the Casimir invariant, with eigenvalues $-k^2\geqslant 0$; the irreducible representation is then labeled by the wave number $k\in\Re$ (the second Casimir, $\vec{L}\cdot\vec{p}$, is zero). We may thus isolate any single "color" k and work with monochromatic optics. On the other hand, the Maxwell fish-eye seen here has the symmetry algebra so(4) and its representation labels are *discrete*, allowed only for discrete colors ν_n, $n=0,1,2,\ldots$. Once one n is chosen, the space of wave functions is of finite dimensions n^2, just as in the hydrogen atom. All these so(4) representations fit into a single degenerate representation of its dynamical algebra so(4,1).

The dynamical algebra so(4,1) also contains the *Euclidean* iso(3) algebra of rotations and space translations of homogeneous media. Homogeneous and fish-eye optical spaces are thus identified as different subalgebra reductions of their common dynamical algebra so(4,1). The iso(3) algebra is a contraction of so(4) by the scaling generator $\Lambda_{N,N+1}$ seen in (5.4d). In this sense, so(4,1) is the *potential* algebra[19] (or group,[20] that binds the fish-eye light orbits to free propagation in a homogeneous optical medium.

The role of the larger dynamical algebra so(4,2) in the Maxwell fish-eye is more subtle: it yields the Hamiltonian time evolution as a *number* generator in the algebra, $\mathcal{N}=\Lambda_{5,6}$, that exponentiates easily to the evolution subgroup. Indeed, the same strategy of finding a larger group, applied to homogeneous space optics, will use the square root p of the Casimir invariant p^2 of iso(3). This quantity is $p=(\Lambda_{4,6}-\Lambda_{5,6})/2\rho$ and commutes with all iso(3) operators. The scaling generator also contracts $\mathcal{N}$ to p.

We have seen here that there is a realization of the algebras that models scalar geometrical optics, and another realization that models *wave* optics, both for the Maxwell fish-eye and for homogeneous media. Evolution along an optical axis has been the primary concern for Euclidean optics,[9] while evolution in time is highlighted here to predict chromatic dispersion.

Small neighborhoods around points q_0 in a smooth radially symmetric inhomogeneous medium may be approximated by neighborhoods of Maxwell fish-eyes that are displaced by $\mathbf{v}$ and/or scaled by ρ and n_0, in such a way as to approximate the refractive index $n(\mathbf{q}_0+\mathbf{q})=n(\mathbf{q}_0)+\Sigma_i q_i[\partial_{qi}n(\mathbf{q})]_{\mathbf{q}_0}+\cdots$ by the "local" fish-eye shape

$$n_0/(1+|\mathbf{q}-\mathbf{v}|^2/4\rho^2)\sim n_0(1+|\mathbf{v}|^2/4\rho^2)-n_0\mathbf{q}\cdot\mathbf{v}/2\rho^2+\cdots$$

through their value and gradient, when $n(|\mathbf{q}|)$. This construction is a section in a bundle over configuration space, where for each $\mathbf{q}\in\Re^3$ there is an so(4,2) evolution direction determined by the local Hamiltonian. While the curvature is positive, the "compact" $\Lambda_{5,6}$-generated subgroup is followed, or its translates by SO(2,1)$\subset$SO(4,2) group transformations. In the regions where n is constant, the direction is along the free-flight Euclidean number operator p. Finally, when the curvature is negative ($\rho\mapsto i\rho$), the "noncompact" generator is $\Lambda_{4,6}$. This 'hyperbolic' Maxwell fish-eye carries its corresponding local so(3,1) symmetry algebra. Work is being done to understand further the Lie algebra and global group properties of particular inhomogeneous optical systems.

ACKNOWLEDGMENTS

For references and useful advice on inhomogeneous media and group theory, we would like to thank Miguel Cervantes (Instituto Nacional de Astrofísica, Optica y Electrónica, Tonantzintla), Bogdan Mielnik (Centro de Investigación y Estudios Avanzados, IPN), Enrique López Moreno (Facultad de Ciencias, UNAM), Thomas H. Seligman (Instituto de Física, UNAM-Cuernavaca), Gerardo Torres del Castillo (Universidad Autónoma de Puebla), and R. López-Peña (ICN-UNAM).

[1] O. Stavroudis, *The Optics of Rays, Wavefronts, and Caustics* (Academic, New York, 1972).
[2] M. Born and E. Wolf, *Principles of Optics* (Pergamon, New York, 1959), Sect. 4.2.2.
[3] R. K. Luneburg, *Mathematical Theory of Optics* (University of California, Berkeley, CA, 1964), Sect. 28.
[4] J. C. Maxwell, Cambridge Dublin Math. J. **8**, 188 (1854).
[5] V. Fock, Z. Phys. **98**, 145 (1935).
[6] V. Bargmann, Z. Phys. **99**, 576 (1936).
[7] H. A. Buchdahl, Am. J. Phys. **46**, 840 (1978).
[8] G. Toraldo di Francia, Opt. Acta **1**, 157 (1954-55).
[9] K. B. Wolf, in *Lie Methods in Optics, II Workshop*, Lecture Notes in Physics, Vol. 352 (Springer-Verlag, Heidelberg, 1989).
[10] V. I. Man'ko and K. B. Wolf, in *Lie Methods in Optics, II Workshop*, Lecture Notes in Physics, Vol. 352 (Springer-Verlag, Heidelberg, 1989).
[11] (a) H. Goldstein, *Classical Mechanics* (Addison-Wesley, Reading, MA 1963), Chaps. 7 and 8; (b) E. C. G. Sudarshan and N. Mukunda, *Classical Dynamics: A Modern Perspective* (Wiley, New York, 1974), Chap. 8; (c) K. Sundermeyer, *Constrained Dynamics*, Lecture Notes in Physics, Vol. 169 (Springer-Verlag, Heidelberg, 1982), Chap. III.
[12] S. Steinberg, in *Lie Methods in Optics*, Lecture Notes in Physics, Vol. 250 (Springer-Verlag, Heidelberg, 1986).
[13] H. A. Buchdahl, *Introduction to Hamiltonian Optics* (Cambridge U.P., Cambridge, England, 1970).
[14] R. Gilmore, *Lie Groups, Lie Algebras, and Some of their Applications* (Wiley–Interscience, New York, 1978).

[15] B. G. Wybourne, *Classical Groups for Physicists* (Wiley, New York, 1974).
[16] K. B. Wolf, Suppl. Nuovo Cimento **5**, 1041 (1967).
[17] P. M. Morse and H. Feshbach, *Methods of Theoretical Physics* (McGraw-Hill, New York, 1953), p. 1210 *et seq.*
[18] L. C. Biedenharn and J. D. Louck, *Angular Momentum in Quantum Physics. Theory and Application*, Encyclopedia of Mathematics, Vol. 8 (Academic, New York, 1981).
[19] A. Frank and K. B. Wolf, Phys. Rev. Lett. **52**, 1737 (1984); A. Frank and K. B. Wolf, J. Math. Phys. **26**, 973 (1985).
[20] Y. Alhassid, F. Gürsey, and F. Iachello, Ann. Phys. (NY) **148**, 346 (1983); Ann. Phys. (NY) **167**, 181 (1986); J. Wu, F. Iachello, and Y. Alhassid, Ann. Phys. (NY) **173**, 68 (1987).
[21] T. Sekiguchi and K. B. Wolf, Am. J. Phys. **55**, 830 (1987).
[22] E. López-Moreno and K. B. Wolf, Rev. Mex. Fís. **35**, 291 (1989).
[23] G. F. Torres del Castillo, Rev. Mex. Fis. **35**, 691 (1989).
[24] K. B. Wolf, in *Group Theory and its Applications*, edited by E. M. Loebl (Academic, New York, 1975), Vol. 3.
[25] H. Flanders, *Differential Forms with Applications to the Physical Sciences* (Academic, New York, 1963).
[26] Reference 15, Sec. 21, p. 297.
[27] A. O. Barut and G. L. Bornzin, J. Math. Phys. **12**, 841 (1971).
[28] C. P. Boyer and K. B. Wolf, J. Math. Phys. **14**, 1853 (1973); K. B. Wolf and C. P. Boyer, J. Math. Phys. **15**, 2096 (1974).
[29] J. M. Souriau, *Structure des Systèmes Dynamiques* (Dunod, Paris, 1970); B. Kostant, *Quantization and Unitary Representations*, Lecture Notes in Mathematics, Vol. 170 (Springer, New York, 1970); J. Śniatycki, *Geometric Quantization and Quantum Mechanics* (Springer, New York, 1980); V. Aldaya, J. A. de Azcárraga, and K. B. Wolf, J. Math. Phys. **25**, 506 (1984).

On Clebsch–Gordan coefficients and matrix elements of representations of the quantum algebra $U_q(su_2)$

V. A. Groza, I. I. Kachurik, and A. U. Klimyk
Institute for Theoretical Physics, Kiev-130, USSR

(Received 8 November 1989; accepted for publication 20 June 1990)

Clebsch–Gordan coefficients and matrix elements of irreducible representations of the quantum algebra $U_q(su_2)$ were considered in several papers. In particular, a few expressions for them were derived. An approach to Clebsch–Gordan coefficients and to matrix elements of representations of $U_q(su_2)$ on the base of the theory of basic hypergeometric functions is given. This approach allows one to obtain q-analogs of all well-known classical expressions for Clebsch–Gordan coefficients (most of them were absent). New symmetry relations, generating functions, and recurrence formulas for Clebsch–Gordan coefficients of $U_q(su_2)$ are obtained. Unlike other papers, Clebsch–Gordan coefficients and matrix elements are considered on the base of minimal theoretical constructions (in fact, without using the notion of a C^* algebra and of a Hopf algebra).

I. INTRODUCTION

Quantum groups and algebras appeared in the quantum method of the inverse scattering problem.[1-4] An independent definition of compact quantum groups, as deformations of Lie groups, was given by Woronowicz.[5,6] The theory of representations of quantum groups and algebras is under development.[7-10] It was shown[11] that a q-analog of the quantum harmonic oscillator is related to quantum groups. Quantum groups and algebras are of great importance for applications in quantum integral systems, in quantum field theory, and in statistical physics.[12-15] The recent papers[16-19] show that representations of quantum groups and algebras are closely connected with basic hypergeometric functions.

Our paper deals with irreducible finite-dimensional representations of the quantum algebra $U_q(su_2)$. They determinate the quantum group $SU_q(2)$. Finite-dimensional representations of the algebra $U_q(su_2)$ are described in Refs. 3 and 4. Representations of the quantum group $SU_q(2)$ are found in Refs. 6 and 17. Kirillov and Reshetikhin[18] have considered Clebsch–Gordan coefficients (CGC's) and Racah coefficients of $U_q(su_2)$. They gave three expressions for CGC's and derived some of their properties. CGC's are also considered in Refs. 19 and 20, where they are expressed in terms of q-Hahn polynomials and of the basic hypergeometric functions ${}_3\varphi_2$.

In Refs. 18 and 20 expressions for CGC's are given without proof. The derivation of Koelink and Koornwinder[19] is based on the results of the paper by Woronowicz[6] and uses the C^* algebra theory. We give a simple derivation of an expression for CGC's with the help of an analog of classical method of highest weights.[21] Then using the results of the theory of basic hypergeometric functions and of the q-calculus we derive several expressions for CGC's. Almost all of them are analogs of the well-known expressions for CGC's of the group $SU(2)$. These expressions are obtained when $q \to 1$.

We derive new symmetry relations for CGC's of $U_q(su_2)$ and give recurrence formulas and generating functions for them. Using CGC's, we derive expressions for the matrix elements d^l_{mn} of representations T^l of $U_q(su_2)$. The part of these matrix elements [up to constants $c(l,m,n,q)$] was found in Ref. 17. Let us note that matrix elements of representations of the quantum group $SU_q(2)$, which are understood in Woronowicz's sense,[6] are given in Refs. 16 and 22. Of course, there are relations between these matrix elements. But, first, our derivation does not use the C^* algebra theory; second, our matrix elements correspond to the fixed basis, for which CGC's are given; and, third, our matrix elements have somewhat another sense.

Our presentation depends strongly on results of the theory of basic hypergeometric functions and of the q-calculus. We give necessary definitions and formulas in Sec. II. Proofs of these results can be found in Refs. 23–25.

In Sec. III, irreducible representations of the algebra $U_q(su_2)$ and their CGC's are introduced. The expressions for CGC's are derived in Sec. IV. We give them in terms of the basic hypergeometric function ${}_3\varphi_2$ and in the form of sums. Symmetry relations for CGC's are obtained in Sec. V. Generating functions and recurrence relations are given in Secs. VI and VII. In Sec. VIII we show the connection of definition of the quantum group $SU_q(2)$ with matrix elements of representations of $U_q(su_2)$. Further information can be found in Refs. 1 and 17. The expressions for these matrix elements are derived in Sec. IX. They are represented as functions of the matrix elements $t_{11}, t_{12}, t_{21}, t_{22}$ of the two-dimensional representation $T^{1/2}$ of $U_q(su_2)$. The results for the classical group $SU(2)$ are obtained when $q \to 1$. Therefore, the results of Sec. IX give a new approach to the Wigner d-functions. In fact, formulas (86)–(89) below are a q-analog of classical Wigner d-functions.

II. BASIC HYPERGEOMETRIC FUNCTIONS AND q-CALCULUS

Let

$$(a;q)_n = \prod_{j=0}^{n-1} (1 - aq^j), \quad (a;q)_0 = 1, \quad a \in C, \ n \in Z_+, \qquad (1)$$

where q is a fixed complex number and Z_+ is the set of nonnegative integers. Basic hypergeometric functions are defined as

 0022-2488/90/122769-12$03.00

$$
{}_{n+1}\varphi_n(a_1,\ldots,a_{n+1};b_1,\ldots,b_n;q,z)
$$
$$
= {}_{n+1}\varphi_n\left(\begin{matrix}a_1,\ldots,a_{n+1}\\ b_1,\ldots b_n\end{matrix}|q,z\right)
$$
$$
= \sum_{m=0}^{\infty} \frac{(a_1;q)_m\cdots(a_{n+1};q)_m}{(b_1;q)_m\cdots(b_n;q)_m}\frac{z^m}{(q;q)_m}. \tag{2}
$$

We shall also use the notation

$$
{}_{n+1}\Phi_n(c_1,\ldots,c_{n+1};d_1,\ldots,d_n;q,z)
$$
$$
= {}_{n+1}\varphi_n(q^{c_1},\ldots,q^{c_{n+1}};q^{d_1},\ldots,q^{d_n};q,z). \tag{3}
$$

If $a_1 = q^{-n}, n\in Z_+$, then series (2) terminates. It is a polynomial. The polynomial

$$
p_n(x;a,b\,|q) = {}_2\varphi_1(q^{-n},abq^{n+1};aq;q,qx) \tag{4}
$$

is called a little q-Jacobi polynomial.[26]

If one considers matrix elements of representations of $U_q(\mathrm{su}_2)$, the relation

$$
{}_2\varphi_1(q^{-n},b;c;q,z)
$$
$$
= (-1)^n q^{-n(n+1)/2}\{(b;q)_n/(c;q)_n\}z^n
$$
$$
\times {}_2\varphi_1(q^{-n},q^{-n+1}c^{-1};q^{-n+1}b^{-1};q,q^{n+1}c/bz) \tag{5}
$$

is useful. We shall also use the relations[27]

$$
{}_3\varphi_2(q^{-n},d,e;f,g;q,q)
$$
$$
= \frac{(f/d;q)_n}{(f;q)_n} d^n {}_3\varphi_2\left(q^{-n}, d, \frac{g}{e}; \frac{dq^{-n+1}}{f}, g; q, \frac{eq}{f}\right), \tag{6}
$$

$$
{}_2\varphi_1\left(a,b;c;q,\frac{cq^n}{a}\right) = \frac{(c/a;q)_n}{(c;q)_n} \tag{7}
$$

Let us note the formula

$$
{}_2\varphi_1(a,b;c;q,z) = \frac{(abz/c;q)_\infty}{(z;q)_\infty} {}_2\varphi_1\left(\frac{c}{a},\frac{c}{b};c;q,\frac{abz}{c}\right), \tag{8}
$$

where $|q|<1$ and $(a;q)_\infty = \lim_{n\to\infty}(a;q)_n$.

If $a\in C$ then $[a]$ will denote the expression

$$
[a] = \frac{q^{a/2}-q^{-a/2}}{q^{1/2}-q^{-1/2}} = q^{-(a-1)/2}\frac{1-q^a}{1-q}. \tag{9}
$$

If $m\in Z_+$ then

$$
[m]! = [1][2]\cdots[m] = \frac{q^{-m(m-1)/4}}{(1-q)^m}(q;q)_m, \quad [0]! = 1, \tag{10}
$$

where $(q;q)_n$ is given by formula (1). For $(q;q)_n$ we have[28]

$$
(q;q)_{N+r} = (q;q)_N(q^{N+1};q)_r, \tag{11}
$$
$$
(q;q)_{N-r} = (-1)^r q^{r(r-2N-1)/4}(q;q)_N(q^{-N};q)_r^{-1}, \tag{12}
$$
$$
(q^{-1};q^{-1})_r = (-1)^r q^{-r(r+1)/2}(q;q)_r. \tag{13}
$$

The expressions $[N\pm r]!$ can be written in terms of $(q^m;q)_r$:

$$
[N+r]! = [N]!q^{-r(r+2N-1)/4}\{(q^{N+1};q)_r/(1-q)^r\}, \tag{14}
$$
$$
[N-r]! = [N]!(-1)^r q^{r(r-2N-3)/4}
$$
$$
\times\{(1-q)^r/(q^{-N};q)_r\}. \tag{15}
$$

Below we shall also use the relations

$$
(q^{N+1};q)_r = [N+r]![N]!^{-1}q^{r(r+2N-1)/4}(1-q)^r, \tag{16}
$$
$$
(q^{-N};q)_r = [N]![N-r]!^{-1}(-1)^r q^{r(r-2N-3)/4}(1-q)^r. \tag{17}
$$

The expression

$$
(1-x)^{(n)} = (1-q^{-1}x)(1-q^{-2}x)\cdots(1-q^{-n}x)
$$
$$
= (q^{-n}x;q)_n
$$
$$
= \sum_{r=0}^{n}\frac{(q^{-n};q)_r}{(q;q)_r}x^r
$$
$$
= \sum_{r=0}^{n}\begin{bmatrix}n\\ r\end{bmatrix}_q (-x)^r q^{r(r-2n-1)/2}
$$
$$
= {}_1\varphi_0(q^{-n};q,x) \tag{18}
$$

is an analog of the binomial formula in the q-calculus. Here,

$$
\begin{bmatrix}n\\ r\end{bmatrix}_q = \frac{(q;q)_n}{(q;q)_r(q;q)_{n-r}} = \frac{[n]!q^{r(n-r)/2}}{[r]![n-r]!}.
$$

It is a q-analog of the binomial coefficient. Usually one sets

$$
(1-x)^{(a)} = {}_1\varphi_0(q^{-a};q,x) \tag{19}
$$

for any $a\in C$. For $(b-x)^{(n)}$ we have

$$
(b-x)^{(n)} = b^n(1-x/b)^{(n)} = \sum_{r=0}^{n}\frac{(q^{-n};q)_r}{(q;q)_r}x^r b^{n-r}. \tag{20}
$$

In the q-calculus the differentiation operator is replaced by the difference operator $\hat{B}_x$, where

$$
\hat{B}_x f(x) = \{f(qx)-f(x)\}/(qx-x).
$$

We have $\hat{B}_x \to d/dx$ when $q\to 1$. The inverse of the operator $\hat{B}_x$ is the q-integration operator. For $0<q<1$ the definite q-integral is given as

$$
\int_0^c f(x)d_q(x) = c(1-q)\sum_{r=0}^{\infty} q^r f(q^r c)
$$
$$
= \sum_{r=0}^{\infty}(x_r - x_{r+1})f(x_r), \tag{21}
$$

where $c>0$ and $x_r = cq^r$. The formula

$$
E_q(x) = \sum_{r=0}^{\infty}\frac{x^r}{[[r]]!}, \tag{22a}
$$

where

$$
[[r]]! = \prod_{j=1}^{r}[[j]] = \prod_{j=1}^{r}\frac{q^j-1}{q-1} = \frac{(q;q)_r}{(1-q)^r},
$$

defines a q-analog of the exponential function. We have $\hat{B}_x E_q(ax) = aE_q(ax)$. The relation

$$
\int_0^{\infty} x^{b-1}E_q(-x)d_q x
$$
$$
\equiv (1-q)\sum_{n=-\infty}^{\infty}\{q^{n(b-1)}E_q(-q^n)\}q^n
$$
$$
= q^{b(b+1)/2}\Gamma_q(b) \tag{22b}
$$

determines q-gamma function. It has the properties

$$\Gamma_q(n) = [[n-1]]!,$$
$$\Gamma_q(b+1) = [[b]]\Gamma_q(b), \quad n \in Z_+, \quad b \in C. \tag{23a}$$

The formula

$$\int_0^1 u^{a-1}(1-qu)^{(c-a-1)} \times {}_n\Phi_{n-1}(d_1,\ldots,d_n;e_1,\ldots,e_{n-1};q,ux)d_q u = \Gamma_q(a)\Gamma_q(c-a)/\Gamma_q(c) \times {}_{n+1}\Phi_n(d_1,\ldots,d_n,a;e_1,\ldots,e_{n-1},c;q,x), \tag{23b}$$

where $\mathrm{Re}\, a > 0$, $\mathrm{Re}(c-a) > 0$, is a q-analog of the well-known formula for usual hypergeometric functions.[28]

The function ${}_2\varphi_1(a,b;c;q,x) = {}_2\varphi_1(a,b;c;x)$ satisfies the relations

$$(1-a){}_2\varphi_1(qa,b;c;x) = {}_2\varphi_1(a,b;c;x) - a{}_2\varphi_1(a,b;c;qx), \tag{24}$$

$$(1-b){}_2\varphi_1(a,qb;c;x) = {}_2\varphi_1(a,b;c;x) - b{}_2\varphi_1(a,b;c;qx), \tag{25}$$

$$(1-cq^{-1}){}_2\varphi_1(a,b;q^{-1}c;x) = {}_2\varphi_1(a,b;c;x) - cq^{-1}{}_2\varphi_1(a,b;c;qx), \tag{26}$$

$$\{x(1-a)(1-b)/(1-c)\}{}_2\varphi_1(qa,qb;qc;x) = {}_2\varphi_1(a,b;c;x) - {}_2\varphi_1(a,b;c;qx) \tag{27}$$

[see formulas (2.9) and (2.10) in Ref. 29]. Excluding appropriate summands from (24)–(26) we obtain relations that can be written as

$$[e-d]{}_2\Phi_1(d,e;f;x) + q^{e/2}[d]{}_2\Phi_1(d+1,e;f;x) - q^{d/2}[e]{}_2\Phi_1(d,e+1;f;x) = 0, \tag{28}$$

$$[e-d][f-1]{}_2\Phi_1(d,e;f-1;x) + [d][f-e-1]{}_2\Phi_1(d+1,e;f;x) - [e][f-d-1]{}_2\Phi_1(d,e+1;f;x) = 0. \tag{29}$$

Excluding ${}_2\varphi_1(a,b;c;x)$ and ${}_2\varphi_1(a,b;c;qx)$ from (24), (25), and (27) and then replacing qa by a, qb by b, we obtain the relation

$${}_2\Phi_1(d,e-1;f;x) - {}_2\Phi_1(d-1,e;f;x) = q^{(d+e-f-2)/2}[e-d][f]^{-1}x{}_2\Phi_1(d,e;f+1;x). \tag{30}$$

III. REPRESENTATIONS OF THE QUANTUM ALGEBRA $U_q(\mathrm{su}_2)$

The associative algebra, generated by the elements H, E_+, E_-, which obey the commutation relations,

$$[H,E_\pm] = \pm E_\pm, [E_+,E_-] = \frac{\sinh hH}{\sinh h/2} = \frac{q^H - q^{-H}}{q^{1/2} - q^{-1/2}},$$

where $q = \exp h$, is called the quantum algebra $U_q(\mathrm{su}_2)$. It is a deformation of the universal enveloping algebra of the classical Lie algebra su(2). The structure of a Hopf algebra is introduced[1] on $U_q(\mathrm{su}_2)$. This structure includes the homomorphism

$$\Delta: U_q(su_2) \to U_q(\mathrm{su}_2) \otimes U_q(\mathrm{su}_2),$$

which acts onto H, E_+, E_- as

$$\Delta(E_\pm) = E_\pm \otimes q^{H/2} + q^{-H/2} \otimes E_\pm, \tag{31a}$$

$$\Delta(H) = H \otimes 1 + 1 \otimes H. \tag{31b}$$

The relation (31a) means that the tensor product for the algebra $U_q(\mathrm{su}_2)$ is not commutative.

Finite-dimensional irreducible representations of $U_q(\mathrm{su}_2)$ are given[4] by integral or by half integral number l. If the representation T^l corresponds to the number l then the carrier space V_l of T^l has the orthonormal basis e_m, $m = -l, -l+1,\ldots,l$, such that

$$E_\pm e_m = ([l \mp m][l \pm m + 1])^{1/2} e_{m\pm 1}, \quad He_m = me_m, \tag{32}$$

where $[n]$ is given by formula (9). The representations T^l of $U_q(\mathrm{su}_2)$ are deformations of the corresponding representations of the classical Lie algebra su(2). It is easy to verify that

$$(E_\pm)^n e_m = \left(\frac{[l \mp m]![l \pm m + n]!}{[l \mp m - n]![l \pm m]!}\right)^{1/2} e_{m\pm n}, \quad |m \pm n| \leqslant l. \tag{33}$$

Finite-dimensional representations of $U_q(\mathrm{su}_2)$ are completely reducible. The tensor product of representations of $U_q(\mathrm{su}_2)$ are defined in accordance with formulas (31a) and (31b). Therefore, $T^l \otimes T^m \neq T^m \otimes T^l$ and for matrix elements we have

$$d^l_{sl} d^m_{uv} \neq d^m_{uv} d^l_{sl}. \tag{34}$$

The last inequality will be more clear in Sec. VIII. If $T^{l_i}(E_\pm) \equiv E^{l_i}_\pm$, $T^{l_i}(H) \equiv H_i$, then it follows from (31a) and (31b) that

$$E^{\otimes}_\pm = E^{l_1}_\pm \otimes q^{H_2/2} + q^{-H_1/2} \otimes E^{l_2}_\pm, \tag{35a}$$

$$H^{\otimes} = H_1 \otimes 1 + 1 \otimes H_2, \tag{35b}$$

where $E^{\otimes}_\pm$ and $H^{\otimes}$ are the operators in the representation $T^{l_1} \otimes T^{l_2}$. It was proved by Jimbo[4] that

$$T^{l_1} \otimes T^{l_2} = \sum_l \oplus T^l, \tag{36a}$$

where summation is the same as in the case of the classical group SU(2) (see, for example, Ref. 21).

Let $\{e_j\}$, $\{e'_k\}$, and $\{e^l_m\}$ be the bases of the carrier spaces of the representations T^{l_1}, T^{l_2}, and T^l from (36a), in which E_+, E_-, H are given by the formulas of the type (32). As in the classical case, we have the expansion

$$e^l_m = \sum_{j,k} \begin{bmatrix} l_1 & l_2 & l \\ j & k & m \end{bmatrix} e_j \otimes e'_k, \tag{36b}$$

which defines CGC's of the tensor product $T^{l_1} \otimes T^{l_2}$. The formula (31b) means that CGC is equal to zero if $j + k \neq m$. Everywhere below we assume that $j + k = m$. Similar to classical case,[21] CGC's of $U_q(\mathrm{su}_2)$ satisfy the orthogonality relations

$$\sum_j \begin{bmatrix} l_1 & l_2 & l \\ j & k & m \end{bmatrix} \begin{bmatrix} l_1 & l_2 & l' \\ j & k & m \end{bmatrix}^* = \delta_{ll'}, \tag{37a}$$

$$\sum_l \begin{bmatrix} l_1 & l_2 & l \\ j & m-j & m \end{bmatrix} \begin{bmatrix} l_1 & l_2 & l \\ j' & m-j' & m \end{bmatrix}^* = \delta_{jj'}. \tag{37b}$$

IV. CLEBSCH–GORDAN COEFFICIENTS OF $U_q(su_2)$

To evaluate CGC's of $U_q(su_2)$ we use the method of highest weights.[21] We put $m = l$ into relation (36b) and act by the operator $T^l(E_+) \equiv E^l{}_+$ upon both sides of this relation. According to (35a) we have

$$E^l{}_+\left(\sum_{j+k=1}\begin{bmatrix} l_1 & l_2 & l \\ j & k & l \end{bmatrix} e_j \otimes e'_k\right)$$
$$= \sum_{j+k=1}\begin{bmatrix} l_1 & l_2 & l \\ j & k & l \end{bmatrix}$$
$$\times (q^{k/2}(E^{l_1}{}_+ e_j) \otimes e'_k + q^{-j/2} e_j \otimes E^{l_2}{}_+ e'_k).$$

In addition, $E^l{}_+ e^l_l = 0$. Using this formula, in the same way as in Ref. 21 we obtain the recurrence relation for CGC's

$$\begin{bmatrix} l_1 & l_2 & l \\ j & k & l \end{bmatrix},$$

which gives,

$$\begin{bmatrix} l_1 & l_2 & l \\ j & k & l \end{bmatrix} = (-1)^{l_1 - j} q^{(-j-1)(l_1-j)/2} \times \left(\frac{[l_1+j]![l_2+k]!}{[l_1-j]![l_2-k]!}\right)^{1/2} A, \tag{38}$$

where A does not depend on j and k. To evaluate A, we substitute expression (38) into formula (37a) with $m = l$. We have

$$A^{-2} = \sum_{j+k=l} q^{(-j-1)(l_1-j)/2} \frac{[l_1+j]![l_2+k]!}{[l_1-j]![l_2-k]!}.$$

Using relations (14) and (15) and the definition of the function ${}_2\Phi_1$ we obtain

$$A^{-2} = \frac{[2l_1]![l+l_2-l_1]!}{[l_1+l_2-l]!} \times {}_2\Phi_1(l+l_2-l_1+1, l-l_1-l_2; -2l_1; q, q^{-2l-1}).$$

Applying formula (7) we represent this basic hypergeometric function in the form of one summand. Now in the same way as in Ref. 21 we get

$$A^{-1} = q^{-(l_1+l_2-l)(l+l_2-l_1+1)/2}\left(\frac{[2l+1]![l_1+l_2-l]!}{[-l_1+l_2+l]![l+l_1-l_2]![l+l_1+l_2+1]!}\right)^{1/2}.$$

Therefore,

$$\begin{bmatrix} l_1 & l_2 & l \\ j & k & l \end{bmatrix} = (-1)^{l_1-j} q^{\{l_2(l_2+1) - l_1(l_1+1) - l(l+1) + 2j(l+1)\}/4} \times \left(\frac{[2l+1]![l_1+l_2-l]![l_1+j]![l_2+k]!}{[l_1+l_2+l+1]![l+l_1-l_2]![l-l_1+l_2]![l_1-j]![l_2-k]!}\right)^{1/2}, \tag{39}$$

where $[n]!$ is defined by formula (10). According to formula (33),

$$(E^l{}_-)^{l-m} e^l_l = \left(\frac{[2l]![1-m]!}{[l+m]!}\right)^{1/2} e^l_m.$$

In this reason

$$\begin{bmatrix} l_1 & l_2 & l \\ j & k & m \end{bmatrix} = \left(\frac{[l+m]!}{[2l]![l-m]!}\right)^{1/2} \langle (E^l{}_-)^{l-m} e^l_l, e_j \otimes e'_k \rangle = \left(\frac{[l+m]!}{[2l]![l-m]!}\right)^{1/2} \langle e^l_l, (E^l{}_+)^{l-m}(e_j \otimes e'_k) \rangle. \tag{40}$$

Using the mathematical induction it is easy to derive that

$$(E^l{}_+)^{l-m} \equiv (E^{l_1}{}_+ \otimes q^{H_2/2} + q^{-H_1/2} \otimes E^{l_2}{}_+)^{l-m} = \sum_{k=0}^{l-m} \begin{bmatrix} l-m \\ k \end{bmatrix}_{q^{-1}} \times (E^{l_1}{}_+ \otimes q^{H_2/2})^k (q^{H_1/2} \otimes E^{l_2}{}_+)^{l-m-k}.$$

Hence, with the help of formulas (10) and (13) we obtain from (40) that

$$\begin{bmatrix} l_1 & l_2 & l \\ j & k & m \end{bmatrix} = \sum_{n=0}^{l-m} q^A \frac{[l-m]!}{[n]![l-m-n]!}\left(\frac{[l_1-j]![l_2-k]!}{[l_1+j]![l_2+k]!}\right. \times \left.\frac{[l_2+l-j-n]![l_1+j+n]!}{[l_2-l+j+n]![l_1-j-n]!}\right)^{1/2} \times \begin{bmatrix} l_1 & l_2 & l \\ j+n & k+l-m-n & l \end{bmatrix},$$

where $A = (nk - j(l-m-n))/2$. From here and from (39) we derive a q-analog of Racah formula for CGC's:

$$\begin{bmatrix} a & b & c \\ j & k & m \end{bmatrix} = (-1)^{a-j} q^{B_1} \times \left(\frac{[c-m]![c+m]![a-j]![b-k]![a+b-c]![2c+1]}{[a+j]![b+k]![a-b+c]![c-a+b]![a+b+c+1]!}\right)^{1/2} \times \sum_n \frac{(-1)^n q^{n(c+m+1)/2}[a+j+n]![b+c-j-n]!}{[n]![c-m-n]![a-j-n]![b-c+j+n]!}, \tag{41a}$$

where $m = j + k$ and

$$B_1 = \tfrac{1}{4}b(b+1) - \tfrac{1}{4}a(a+1) - \tfrac{1}{4}c(c+1) + \tfrac{1}{2}j(m+1)$$

(we replace l_1, l_2, and l by a, b, and c, respectively). Let us note that this formula is given without proof in Refs. 18 (with a misprint) and 20.

By means of relations (14) and (15) the sum in (41a) is reduced to the form (2) for the function ${}_3\varphi_2$. Replacing ${}_3\varphi_2$ by ${}_3\Phi_2$ [see formula (3)], we receive

$$\begin{bmatrix} a & b & c \\ j & k & m \end{bmatrix} = \frac{(-1)^{a-j} q^{B_1} \Delta(abc)\,[c+b-j]!}{[a-b+c]![c+b-a]![b-c+j]!} \times \left(\frac{[a+j]![b-k]![c+m]![2c+1]}{[a-j]![b+k]![c-m]!} \right)^{1/2} {}_3\Phi_2\left(\begin{matrix} -a+j, a+j+1, -c+m \\ b-c+j+1, -b-c+j \end{matrix} \middle| \, q,q \right), \tag{41b}$$

where B_1 is the same as in (41a) and

$$\Delta(abc) = \left(\frac{[a+b-c]![a-b+c]![c-a+b]!}{[a+b+c+1]!} \right)^{1/2}.$$

The relation (6) is used to obtain other expressions for CGC's of $U_q(\mathrm{su}_2)$. Setting

$$n = a - j, \quad d = q^{a+j+1}, \quad e = q^{m-c},$$
$$f = q^{j-c+b+1}, \quad g = q^{j-c-b}$$

into (41b) and applying relation (6), we get

$$\begin{bmatrix} a & b & c \\ j & k & m \end{bmatrix} = q^{B_2} \frac{\Delta(abc)\,[b+c-j]!}{[c-b+j]![a+b-c]![b-a+c]!} \times \left(\frac{[a+j]![b-k]![c+m]![2c+1]}{[a-j]![b+k]![c-m]!} \right)^{1/2} {}_3\Phi_2\left(\begin{matrix} a+j+1, -a+j, -b-k \\ -b-c+j, c-b+j+1 \end{matrix} \middle| \, q, q^{-b+k} \right), \tag{42}$$

where

$$B_2 = \tfrac{1}{4}a(a+1) + \tfrac{1}{4}(b+1) - \tfrac{1}{4}c(c+1) + \tfrac{1}{2}jk.$$

Putting here

$$n = a - j, \quad d = q^{-b-k}, \quad e = q^{c-a-b},$$
$$f = q^{c-a-k+1}, \quad g = q^{c-b+j+1},$$

with the help of relation (6) we obtain

$$\begin{bmatrix} a & b & c \\ j & k & m \end{bmatrix} = q^{B_3} \frac{\Delta(abc)}{[a+b-c]![c-a-k]![c-b+j]!} \times ([a+j]![b-k]![c-m]![2c+1][c+m]!)^{1/2} \times ([a-j]!^{-1}[b+k]!^{-1})^{1/2} \times {}_3\Phi_2\left(\begin{matrix} c-a-b, & -b-k, -a+j \\ c-a-k+1, & c-b+j+1 \end{matrix} \middle| \, q,q \right), \tag{43}$$

where

$$B_3 = \tfrac{1}{4}(a+b-c)(a+b+c+1) + \tfrac{1}{2}(ak - bj).$$

Putting

$$n = a - j, \quad d = q^{m-c}, \quad e = q^{a+j+1},$$
$$f = q^{b-c+j+1}, \quad g = q^{j-b-c},$$

into (41a) and using the relation (6) we obtain

$$\begin{bmatrix} a & b & c \\ j & k & m \end{bmatrix} = (-1)^{a-j} q^{B_4} \frac{[b+c-j]![a+b-m]!}{\Delta(abc)\,[a+b+c+1]!} \times \left(\frac{[a+j]![c+m]![2c+1]}{[a-j]![b-k]![b+k]![c-m]!} \right)^{1/2} \times {}_3\Phi_2\left(\begin{matrix} -a+j, -a-b-c-1, -c+m \\ -a-b+m, -b-c+j \end{matrix} \middle| \, q, q^{a-b+c+1} \right), \tag{44}$$

where

$$B_4 = \tfrac{1}{4}(b-a-c)(a+b+c+1) + \tfrac{1}{2}j(b+c+1) + \tfrac{1}{2}ak.$$

Let us here set

$$n = a + b + c + 1, \quad d = q^{m-c}, \quad e = q^{a-b-c},$$
$$f = q^{-2c}, \quad g = q^{j-b-c},$$

and apply relation (6), then put

$$n = b - a + c, \quad d = q^{-a-b-c-1},$$
$$e = q^{m-c}, \quad f = q^{j-b-c}, \quad g = q^{-2c},$$

into the obtained formula and again apply this relation. As a result we have

$$\begin{bmatrix} a & b & c \\ j & k & m \end{bmatrix}$$
$$= (-1)^{b+k} q^{B_5} \frac{\Delta(abc)\,[2c]![b+c+j]!}{[a-b+c]![b-a+c]!}$$
$$\times \left(\frac{[a-j]!}{[a+j]![b+k]![b-k]!} \frac{[2c+1]}{[c+m]![c-m]!} \right)^{1/2} {}_3\Phi_2\left(\begin{matrix} a-b-c, -c-m, -a-b-c-1 \\ -2c, -b-c-j \end{matrix} \middle| \, q, q^{b+k+1} \right), \quad (45)$$

where

$$B_5 = \tfrac{1}{4}(a-b-c)(a+b+c+1) - \tfrac{1}{2}k(b+c+1) - \tfrac{1}{2}bj.$$

If we here put

$$n = b-a+c, \quad d = q^{-c-m}, \quad e = q^{a-j+1},$$
$$f = q^{a-b-m+1}, \quad g = q^{-b-c-j},$$

and take into account relation (6), then we obtain

$$\begin{bmatrix} a & b & c \\ j & k & m \end{bmatrix}$$
$$= (-1)^{b+k} q^{B_6} \frac{\Delta(abc)\,[b+c+j]!}{[b-a+c]![a-b-m]!}$$
$$\times \left(\frac{[a-j]![c-m]![2c+1]}{[a+j]![b+k]![b-k]![c+m]!} \right)^{1/2}$$
$$\times {}_3\Phi_2\left(\begin{matrix} a-b-c, -c-m, a-j+1 \\ -b-c-j, a-b-m+1 \end{matrix} \middle| \, q, q \right), \quad (46)$$

where

$$B_6 = \tfrac{1}{4}(a-b-c)(a+b-c+1) + \tfrac{1}{2}j(c-a) - \tfrac{1}{2}k(1+1).$$

Using formulas (2), (3), (16), and (17) we express the function ${}_3\Phi_2$ in the form of a sum and obtain from (43) a q-analog of the Van der Waerden formula for CGC's:

$$\begin{bmatrix} a & b & c \\ j & k & m \end{bmatrix} = q^{B_5} \Delta(abc) \left([a-j]![a+j]![b+k]![b-k]![c+m]![c-m]![2c+1] \right)^{1/2}$$
$$\times \sum_n \frac{(-1)^n q^{-n(a+b+c+1)/2}}{[n]![a+b-c-n]![a-j-n]![b+k-n]![c-b+j+n]![c-a-k+n]!}. \quad (47)$$

Analogously, a q-analog of the Jusys–Bandzaitis formula for CGC's is derived from formula (44):

$$\begin{bmatrix} a & b & c \\ j & k & m \end{bmatrix} = \frac{(-1)^{a-j} q^{B_4}}{\Delta(abc)} \left(\frac{[a+j]![a-j]![c+m]![c-m]![2c+1]}{[b+k]![b-k]!} \right)^{1/2}$$
$$\times \sum_n \frac{(-1)^n q^{n(a-b+c)/2} [a+b-m-n]![b+c-j-n]!}{[n]![c-m-n]![a-j-n]![a+b+c-n+1]!}. \quad (48)$$

The formula (46) leads to a q-analog of the Wigner formula for CGC's:

$$\begin{bmatrix} a & b & c \\ j & k & m \end{bmatrix} = (-1)^{b+k} q^{B_6} \Delta(abc) \left(\frac{[c+m]![c-m]![2c+1]}{[a+j]![a-j]![b+k]![b-k]!} \right)^{1/2}$$
$$\times \sum_n \frac{(-1)^n q^{n(a+b-c+1)/2} [a-j+n]![b+c+j-n]!}{[n]![c+m-n]![b-a+c-n]![a-b-m+n]!}. \quad (49)$$

Setting $c = a + b$ into (47) we obtain

$$\begin{bmatrix} a & b & a+b \\ j & k & m \end{bmatrix} = q^{(ak-bj)/2}$$
$$\times \left(\frac{[2a]![2b]![a+b-m]![a+b+m]!}{[2a+2b]![a+j]![a-j]![b+k]![b-k]!} \right)^{1/2}. \quad (50)$$

If $c = a - b$ in (49) then we obtain

$$\begin{bmatrix} a & b & a-b \\ j & k & m \end{bmatrix} = (-1)^{b+k} q^{-(bj+ak+k)/2}$$
$$\times \left(\frac{[2b]![2a-2b+1]![a+j]![a-j]!}{[2a+1]![b+k]![b-k]![a-b+m]![a-b-m]!} \right)^{1/2}. \quad (51)$$

V. SYMMETRIES OF CLEBSCH–GORDAN COEFFICIENTS

It follows from (43) that

$$\begin{bmatrix} a & b & c \\ j & k & m \end{bmatrix} = \begin{bmatrix} b & a & c \\ -k & -j & -m \end{bmatrix}. \tag{52}$$

From formula (44) we derive that

$$\begin{bmatrix} a & b & c \\ j & k & m \end{bmatrix} = (-1)^{c-a-k} q^{-k/2} \times \left(\frac{[2c+1]}{[2a+1]}\right)^{1/2} \begin{bmatrix} c & b & a \\ m & -k & j \end{bmatrix}. \tag{53}$$

The expression (43) for CGC's is symmetric with respect to the transformation

$$\begin{bmatrix} a & b & c \\ j & k & m \end{bmatrix} = \begin{bmatrix} (a+b+m)/2 & (a+b-m)/2 & c \\ (a-b+j-k)/2 & (a-b-j+k)/2 & a-b \end{bmatrix}. \tag{54}$$

By means of relation (13) we derive from (41a) that

$$\begin{bmatrix} a & b & c \\ j & k & m \end{bmatrix}_q = (-1)^{a+b-c} \begin{bmatrix} a & b & c \\ -j & -k & -m \end{bmatrix}_q, \tag{55}$$

(since q is replaced by q^{-1} we have supplied CGC's with index q).

The classical analog of symmetry relations (52)–(55) (they correspond to the case $q = 1$) generates the symmetry group for CGC's, which contains 72 elements.[21] In the same way, relations (52)–(55) generate the symmetry group for CGC's of the quantum algebra, containing 72 elements. But now relation (55) replaces q by q^{-1}. In fact, this replacement means that algebra $U_q(\mathrm{su}_2)$ transforms into the algebra $U_{q^{-1}}(\mathrm{su}_2)$.

Besides these symmetries there are "reflection symmetries" of expressions for CGC's related to transition of the type $l \to \bar{l} = -l-1$. As in the classical case, the existence of these symmetries is connected with the fact, that the finite-dimensional representations T^l of $U_q(\mathrm{su}_2)$ are continued [with the help of $U_q(\mathrm{sl}_2)$] to the finite-dimensional representations T^l of $U_q(\mathrm{su}_{1,1})$, which are contained in the reducible representations T^{ϵ}_l and T^{ϵ}_{-l-1} of $U_q(\mathrm{su}_{1,1})$, (for details see Ref. 30).

In the classical case the "reflection symmetries" are related to the property $\Gamma(b+1) = b\Gamma(b)$ of the classical gamma function. The gamma function, introduced in the formula (22b), has analogous property. Let us take the function

$$\hat{\Gamma}_q(b) = q^{-(b-1)(b-2)/4} \Gamma_q(b)$$

instead of $\Gamma_q(b)$. Then according to (23a) we have

$$\hat{\Gamma}_q(n) = [n-1]!,$$
$$\hat{\Gamma}_q(b+1) = [b]\hat{\Gamma}_q(b), \quad n \in Z_+, b \in C.$$

Therefore, if b is not an integer, then

$$\frac{\hat{\Gamma}_q(b)}{\hat{\Gamma}_q(b-n)} = [b-1][b-2]\cdots[b-n]$$
$$= (-1)^n \{\hat{\Gamma}_q(-b+n+1)/\hat{\Gamma}_q(-b+1)\}.$$

If $b \to m$, $m \in Z$, we obtain

$$\frac{\hat{\Gamma}_q(-k+1)}{\hat{\Gamma}_q(-m+1)} = \frac{[-k]!}{[-m]!} = (-1)^{k-m} \frac{[m-1]!}{[k-1]!}. \tag{56}$$

Thus, as in the classical case, the ratio of the values of the gamma function $\hat{\Gamma}_q$ at integer points has sense.

The relations for CGC's, related to "reflection symmetries," are obtained with the help of the expressions for CGC's (in the form of a sum) by substitutions of the type $l \to -l-1 \equiv \bar{l}$ and by taking into account relation (56). In this way, repeating reasonings of the classical case, we obtain the symmetries

$$\begin{bmatrix} a & b & c \\ j & k & m \end{bmatrix} = (-1)^{b+k} \begin{bmatrix} \bar{a} & b & \bar{c} \\ j & k & m \end{bmatrix}$$
$$= (-1)^{b-c-j} \begin{bmatrix} \bar{a} & b & c \\ j & k & m \end{bmatrix}$$
$$= (-1)^{a-j} \begin{bmatrix} a & \bar{b} & \bar{c} \\ j & k & m \end{bmatrix}$$
$$= (-1)^{a+b-c} \begin{bmatrix} \bar{a} & \bar{b} & \bar{c} \\ j & k & m \end{bmatrix}. \tag{57}$$

We can combine these symmetries with symmetries (52)–(55) as it is done in the classical case.

Let us note that symmetry relations (52)–(55) were obtained by other (more complicated) methods in Refs. 18 and 19. The reflection symmetries are new.

VI. GENERATING FUNCTIONS FOR CLEBSCH–GORDAN COEFFICIENTS

Let

$$F = (1+x)^{(b-a+c)} \times {}_2\Phi_1(c-a-b, j-a; c-b+1; q, -q^{a-b-c}x),$$

where $(1+x)^{(n)}$ is defined by formulas (19) and (20). We represent $(1+x)^{(b-a+c)}$ in the form of the sum over n [according to formula (18)] and ${}_2\Phi_1(\cdots)$ in the form of the sum over m [according to formula (2)]. Now we set $m+n = b+k$ and change the summations over n and m by the summations over k and m. After some simple transformations we obtain

$$F = \sum_k \frac{(q^{a-b-c};q)_{b+k}}{(q;q)_{b+k}} x^{b+k} \times {}_3\Phi_2\left(\begin{matrix} c-a-b, j-a, -b-k \\ c-b+j+1, c-a-k+1 \end{matrix} \middle| \, q, q\right).$$

Using formula (43) for CGC's and relation (10) we derive that

$$\left(\frac{[a-b+c]![a+j]![2c+1]}{[b+c-a]![a+b-c]![a+b+c+1]![a-j]!}\right)^{1/2} \times \frac{(1+x)^{(b+c-a)}}{[c-b+j]!} \times {}_2\Phi_1(c-b-a, j-a; c-b+j+1; q, -q^{a-b-c}x)$$

$$= \sum_{k-m=j} \begin{bmatrix} a & b & c \\ j & k & m \end{bmatrix} \times \frac{q^A x^{b+k}}{([b+k]![b-k]![c+m]![c-m]!)^{1/2}}, \quad (58)$$

where

$$A = \tfrac{1}{4}b(2a-2j-2c-3b-3) - (a-c)(a+c+1).$$

Thus, the function from the left-hand side of (58) is a generating function for CGC's of $U_q(\mathrm{su}_2)$. Other similar formulas can be received by means of symmetry relations.

Let us consider formula (58) as an expansion of the function from the left-hand side into the series in powers of x. Then we get

$$\begin{bmatrix} a & b & c \\ j & k & m \end{bmatrix} = M \frac{d^{b+k}}{dx^{b+k}} \{(1+x)^{(b+c-a)}{}_2\Phi_1(c-b-a,j-a;c-b+j+1;q,-q^{a-b-c}x)\}\Big|_{x=0},$$

where

$$M = \frac{q^{-A}}{(b+k)!}\left(\frac{[b-k]![b+k]![c+m]![c-m]![a-b+c]![a+j]![2c+1]}{[b+c-a]![a+b-c]![a+b+c+1]![a-j]![c-b+j]!^2}\right)^{1/2}.$$

There is the relation that can be considered (in some sense) as the inverse of (58). To obtain this relation we need one additional expressions for CGC's. We represent the expression (45) for CGC's in the form of the sum over n and change this summation by the summation over $s = l + m - n$. Then we represent the obtained expression in terms of the function ${}_3\Phi_2$. As a result we have

$$\begin{bmatrix} a & b & c \\ j & k & m \end{bmatrix} = (-1)^{b+c+k+m} q^A \times \frac{\Delta(abc)[a+b+c+1]!}{[a-b+c]![a+b-m+1]![b-a-m]!} \times ([a-j]![b-k]![c-m]![2c+1])^{1/2}([a+j]![b+k]![c+m]!)^{1/2} \times {}_3\Phi_2(-c-m,c-m+1,b-k+1;a+b-m+2,b-a-m+1;q,q), \quad (59a)$$

where

$$A = \tfrac{1}{4}a(a+1) - \tfrac{1}{4}b(b+1) + \tfrac{1}{4}c(c+1)\tfrac{1}{2}k(m-1).$$

Applying the formula (23b) to this function ${}_3\Phi_2$ and taking into account relation (21), we get

$$D\begin{bmatrix} a & b & c \\ j & k & m \end{bmatrix} = \int_0^1 u^{b-k}(1-qu)^{(a-j)}{}_2\Phi_1(-c-m,c-m+1;b-a-m+1;q,u)d_q u$$

$$= (1-q)\sum_{r=0} q^{r(b-k+1)}(1-q^{r+1})^{(a-j)}{}_2\Phi_1(-c-m,c-m+1;b-a-m+1;q,q^r), \quad (59b)$$

where

$$D = \frac{(-1)^s}{q^A}\frac{[a-b+c]![a+b-m+1]!([a+j]![a-j]![b+k]![b-k]![c+m]!)^{1/2}}{\Delta(abc)[a+b+c+1]!([c-m]![2c+1])^{1/2}},$$

$s = b + c + k + m$ and A is the same as in (59a).

Let us show that the function

$$G(x,y,z) = (x-qy)^{(a+b-c)} \times (y-z)^{(b+c-a)}(z-x)^{(a-b+c)} \quad (60a)$$

is also a generating function for CGC's. Representing every multiplier in the form of the sum (20), we obtain

$$G(x,y,z) = \sum_{n=0}^{a+b-c}\sum_{r=0}^{b+c-a}\sum_{s=0}^{a-b+c} \frac{(q^{c-a-b};q)_n(q^{a-b-c};q)_r}{(q;q)_n(q;q)_r(q;q)_s} \times (q^{b-a-c};q)_s q^n x^{s-n+a+b-c} \times y^{n-r+b+c-a} z^{r-s+a-b+c}.$$

Now we put

$$s-n+a+b-c = a+j, \quad n-r+b+c-a = b+k, \quad r-s+a-b+c = c-m$$

and change the summations over n, r, s by the summations over n, j, k. Then

$$G(x,y,z)=\sum_{j,k} x^{a+j}y^{b+k}z^{c-m}\frac{(q^{a-b-c};q)_{c-a-k}(q^{b-a-c};q)_{c-b+j}}{(q;q)_{c-a-k}(q;q)_{c-b+j}}$$
$$\times{}_3\Phi_2(c-a-b,-b-k,j-a;c-a-k+1,c-b+j+1;q,q).$$

Taking into account expression (43) for CGC's, after some simple transformations we receive

$$G(x,y,z)=(-1)^{a+b-2c}\frac{\Delta(abc)[a+b+c+1]!}{[2c+1]^{1/2}}$$
$$\times\sum_{j,k}(-1)^{k-j}$$
$$\times\frac{x^{a+j}y^{b+k}z^{c-m}q^D}{I(a,b,c,j,k)}\begin{bmatrix}a & b & c\\ j & k & m\end{bmatrix}, \tag{60b}$$

where $j+k=m$ and

$$I(a,b,c,j,k)=([a+j]![a-j]![b+k]!$$
$$\times[b-k]![c+m]![c-m]!)^{1/2}, \tag{61}$$

$$D=\tfrac{1}{2}(c-a-k)(a-b-c-1)+\tfrac{1}{2}(c-b+j)$$
$$\times(b-c-a-1)-\tfrac{1}{4}(a+b-c)(a+b+c+1)$$
$$+\tfrac{1}{2}(bj-ak).$$

Next generating function is given by the formula

$$(v_1u_2-qu_1v_2)^{(a+b-c)}(w_1v_2-w_2v_1)^{(b+c-a)}$$
$$\times(u_1w_2-w_1u_2)^{(a-b+c)}$$
$$=\Delta(abc)[a+b+c+1]![2c+1]^{-1/2}$$
$$\times\sum_{j,k}\frac{u_1^{a-j}u_2^{a+j}v_1^{b-k}v_2^{b+k}w_1^{c-m}w_2^{c+m}q^D}{I(a,b,c,j,k)}$$
$$\times\begin{bmatrix}a & b & c\\ j & k & m\end{bmatrix}, \tag{62}$$

where D and $I(a,b,c,j,k)$ are the same as in (60b). To prove this formula it is sufficient to represent the left-hand side in the form

$$u_1^{2a}v_1^{2b}w_1^{2c}\left(\frac{u_2}{u_1}-q\frac{v_2}{v_1}\right)^{(a+b-c)}$$
$$\times\left(\frac{v_2}{v_1}-\frac{w_2}{w_1}\right)^{(b+c-a)}\left(\frac{w_2}{w_1}-\frac{u_2}{u_1}\right)^{(a-b+c)}$$

and to apply formula (60b).

Let us replace in (60a) x by $q^{2/3}x$ and z by $q^{1/3}z$ and take into account relation (20). We obtain that formula (60b) will be correct if we replace the left-hand side by

$$(x-q^{1/3}y)^{(a+b-c)}(y-q^{1/3}z)^{(b+c-a)}$$
$$\times(z-q^{1/3}x)^{(a-b+c)}$$

and multiply the expression under the summation sign on the right-hand side by $q^{(2c-a-b+2j-m)/3}$. The corresponding statement is valid for relation (62).

The relation

$$\frac{E_q(xv)}{E_q(xt)}\frac{E_q(yt)}{E_q(yu)}\frac{E_q(zu)}{E_q(qzv)}$$
$$=\sum_{a,b,c,j,k}\begin{bmatrix}a & b & c\\ j & k & m\end{bmatrix}$$
$$\times\frac{x^{b+c-a}y^{a-b+c}z^{a+b-c}u^{a+j}v^{b+k}t^{c-m}(1-q)^Aq^B}{\Delta(abc)I(a,b,c,j,k)} \tag{63}$$

where $m=j+k$, $A=a+b-2c-j+k$, $I(a,b,c,j,k)$ is given by formula (61) and

$$B=-\tfrac{1}{2}a(3a-1)-\tfrac{1}{2}b(3b+1)$$
$$-\tfrac{1}{2}c(3c+1)-\tfrac{1}{2}j(a+c-2b+1)$$
$$+\tfrac{1}{2}k(b+c-2a+1)+ab+ac+bc,$$

is proved in the same way as relation (60b).

VII. RECURRENCE RELATIONS FOR CLEBSCH–GORDAN COEFFICIENTS

The formula

$$Q_n(q^{-x})=Q_n(q^{-x};d,f,N|q)$$
$$={}_3\varphi_2\left(\begin{matrix}q^{-n},dfq^{n+1},q^{-x}\\ dq,q^{-N}\end{matrix}\,\middle|\,q,q\right),\quad n=0,1,\ldots,N,$$

defines q-Hahn polynomials, which are orthogonal on the set $x\in\{0,1,\ldots,N\}$ [see formulas (3.22) and (3.24) in Ref. 31]. These polynomials are special cases of so-called q-Racah polynomials $R_n(\mu(x))$ from Ref. 32. The recurrence relations for $R_n(\mu(x))$, derived in Ref. 32, leads to that for polynomials $Q_n(q^{-x})$. It has the form

$$A_nQ_{n+1}(q^{-x})-\{A_n+C_n-(1-q^{-x})\}Q_n(q^{-x})$$
$$+C_nQ_{n-1}(q^{-x})=0, \tag{64}$$

where

$$A_n=-\frac{(1-dfq^{n+1})(1-dq^{n+1})(1-q^{n-N})}{(1-dfq^{2n+1})(1-dfq^{2n+2})},$$
$$C_n=-\frac{(1-q^n)(1-fq^n)(q^{-N-1}-dfq^n)dq^{n+1}}{(1-dfq^{2n})(1-dfq^{2n+2})}.$$

From the other side, putting

$$c-m=x,\quad a-j=n,\quad q^{b-c+j}=d,$$
$$q^{c-b+j}=f,\quad c+b-j=N$$

into (41b), we express CGC's in terms of q-Hahn polynomials:

$$\begin{bmatrix}a & b & c\\ j & k & m\end{bmatrix}=(-1)^{a-j}q^B\frac{([a+j]![b-k]![c+m]![2c+1])^{1/2}\Delta(abc)[b+c-j]!}{([a-j]![b+k]![c-m]!)^{1/2}[a-b+c]![c+b-a]![b-c+j]!}$$
$$\times Q_{a-j}(q^{m-l};q^{b-c+j},q^{c-b+j},b+c-j|q).$$

Substituting into (64) the expression for q-Hahn polynomials in terms of CGC's, we get the recurrence relation for CGC's:

$$B_a\begin{bmatrix} a+1 & b & c \\ j & k & m \end{bmatrix} - D_a\begin{bmatrix} a & b & c \\ j & k & m \end{bmatrix} + E_a\begin{bmatrix} a-1 & b & c \\ j & k & m \end{bmatrix} = 0, \tag{65}$$

where

$$B_a = [2a]([a-b+c+1][a+b+c+2][a-j+1][a+j+1][a+b-c+1][b+c-a])^{1/2},$$

$$E_a = [2a+2]([b+c-a+1][a+b-c][a+j][a-j][a-b+c][a+b+c+1])^{1/2},$$

$$D_a = q^{(c+k)/2}[2a][2a+1][2a+2][c-m] - q^{-(a+1)/2}[2a][a+j+1][a+b-c+1][b+c-a] - q^{a/2} \times [2a+2][a-j][a+c-b][a+b+c+1].$$

Using formulas (29) and (2.10) of Ref. 29, we derive the recurrence relations

$$([b-a+c][a-b+c+1])^{1/2}\begin{bmatrix} a & b & c \\ j & k & m \end{bmatrix} = q^{(a-b-j+k)/4}([b+k][a+j+1])^{1/2} \times \begin{bmatrix} a+1/2 & b-1/2 & c \\ j+1/2 & k-1/2 & m \end{bmatrix} + q^{-(a-b+j-k)/4}([b-k][a-j+1])^{1/2} \times \begin{bmatrix} a+1/2 & b-1/2 & c \\ j-1/2 & k+1/2 & m \end{bmatrix},$$

$$([a+b-c][c+m+1][2c+2])^{1/2}\begin{bmatrix} a & b & c \\ j & k & m \end{bmatrix} = q^{(b+c-2a+j+1)/4}([a-b+c+1][b-k])^{1/2} \times \begin{bmatrix} a & b-1/2 & c+1/2 \\ j & k+1/2 & m+1/2 \end{bmatrix} - q^{-(a+c-2b+k+1)/4}([b-a+c+1][a-j])^{1/2} \times \begin{bmatrix} a-1/2 & b & c+1/2 \\ j+1/2 & k & m+1/2 \end{bmatrix},$$

$$([a+b+c+1][c-m][2c][2c+1]^{-1})^{1/2}\begin{bmatrix} a & b & c \\ j & k & m \end{bmatrix} = q^{(c-a+2b-k)/4}([a-j][a-b+c])^{1/2} \times \begin{bmatrix} a-1/2 & b & c-1/2 \\ j+1/2 & k & m+1/2 \end{bmatrix} + q^{(c-b+2a-j)/4}([b-k][b-a+c])^{1/2} \times \begin{bmatrix} a & b-1/2 & c-1/2 \\ j & k+1/2 & m+1/2 \end{bmatrix}.$$

Derivation of these formulas and other recurrence relations will be given in a separate paper.

In the conclusion of this section we give the expressions for CGC's of the tensor product $T^l \otimes T^{1/2}$, which will be used below. They follow from formulas (50) and (51) and have the form

$$\begin{bmatrix} l & 1/2 & l+1/2 \\ j & \pm 1/2 & m \end{bmatrix} = q^{\pm(l\mp m+1/2)/4}\left(\frac{[l\pm m+1/2]}{[2l+1]}\right)^{1/2}, \tag{66a}$$

$$\begin{bmatrix} l & 1/2 & l-1/2 \\ j & \pm 1/2 & m \end{bmatrix} = \mp q^{\mp(l\pm m+1/2)/4}\left(\frac{[l\mp m+1/2]}{[2l+1]}\right)^{1/2}, \tag{66b}$$

where $m = j \pm 1/2$. Let us note that CGC's of these tensor product for somewhat other definition of the quantum algebra $U_q(\mathrm{su}_2)$ are given by Pasquier.[33] His CGC's are obtained from ours if one changes q by q^{-1}.

VIII. QUANTUM GROUP $SU_q(2)$

In the basis e_m, $m = -l, -l+1, \ldots, l$, the representation T^l of $U_q(\mathrm{su}_2)$ is given by the matrix elements d^l_{mn} which depend on elements $a \in U_q(\mathrm{su}_2)$. As in the classical case, the relations

$$\sum_{l}\begin{bmatrix} l_1 & l_2 & l \\ j_1 & k_1 & m_1 \end{bmatrix}\begin{bmatrix} l_1 & l_2 & l \\ j_2 & k_2 & m_2 \end{bmatrix} \times d^l_{m_1 m_2} = d^l_{j_1 j_2} d^{l_2}_{k_1 k_2}, \tag{67a}$$

$$\sum_{j_1 j_2 k_1 k_2}\begin{bmatrix} l_1 & l_2 & l \\ j_1 & k_1 & m_1 \end{bmatrix}\begin{bmatrix} l_1 & l_2 & l \\ j_2 & k_2 & m_2 \end{bmatrix} \times d^{l_1}_{j_1 j_2} d^{l_2}_{k_1 k_2} = d^l_{m_1 m_2} \tag{67b}$$

are valid. We assume that CGC's are real.

The matrix elements $d^{1/2}_{1/2,1/2}$, $d^{1/2}_{1/2,-1/2}$, $d^{1/2}_{-1/2,1/2}$, $d^{1/2}_{-1/2,-1/2}$ of the two-dimensional representations $T^{1/2}$ of $U_q(\mathrm{su}_2)$ will be denoted by t_{11}, t_{12}, t_{21}, t_{22}, respectively. It follows from (66a) and (66b) that CGC's of the tensor product $T^{1/2} \otimes T^{1/2}$ are

$$\begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 1 \\ \frac{1}{2} & \frac{1}{2} & 1 \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 1 \\ -\frac{1}{2} & -\frac{1}{2} & -1 \end{bmatrix} = 1,$$

$$\begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 1 \\ \frac{1}{2} & -\frac{1}{2} & 0 \end{bmatrix} = -\begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ -\frac{1}{2} & \frac{1}{2} & 0 \end{bmatrix} = q^{-1/4}[2]^{-1/2},$$

$$\begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 1 \\ -\frac{1}{2} & \frac{1}{2} & 0 \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & -\frac{1}{2} & 0 \end{bmatrix} = q^{1/4}[2]^{-1/2}.$$

Putting $l_1 = l_2 = 1/2$ into (67a) and using these CGC's, we find that

$$t_{22}t_{11} = q^{1/2}[2]^{-1}(d^{1}_{00} + q^{-1}I),$$
$$t_{11}t_{22} = q^{-1/2}[2]^{-1}(d^{1}_{00} + qI).$$

where $I = d^{0}_{00}$. Hence,

$$t_{22}t_{11} - qt_{11}t_{22} = (q-1)I. \tag{68}$$

In the same way we obtain that

$$t_{12}t_{21} = t_{21}t_{12}, \quad t_{12}t_{11} = \sqrt{q}t_{11}t_{12}, \quad t_{21}t_{11} = \sqrt{q}t_{11}t_{21}, \tag{69}$$

$$\sqrt{q}t_{11}t_{22} - t_{21}t_{12} = \sqrt{q}I, \quad t_{22}t_{12} = \sqrt{q}t_{12}t_{22}, \tag{70}$$

$$t_{22}t_{21} = \sqrt{q}t_{21}t_{22},$$

$$t_{22}t_{11} - t_{11}t_{22} = (\sqrt{q} - 1/\sqrt{q})\, t_{21}t_{12}. \tag{71}$$

Using formula (36a) it is easy to show that any finite-dimensional irreducible representation of $U_q(\mathrm{su}_2)$ can be received by successive tensor multiplication of representations $T^{1/2}$. If in the multiplication procedure we use CGC's, then we obtain matrix elements of the representations T^l, $l = 0,1/2,1,3/2,\ldots$. Therefore, taking products of the elements t_{11}, t_{12}, t_{21}, t_{22} and their linear combinations we get the algebra A which contain all matrix elements of the representations T^l of the algebra $U_q(\mathrm{su}_2)$. The algebra A is an associative algebra, generated by the elements $I = d^{0}_{00}$, t_{11}, t_{12}, t_{21}, t_{22} and by the relations[1,17] (68–71). The structure of the Hopf algebra is introduced into A. In this case A is called the quantum group $\mathrm{SU}_q(2)$. the matrix elements d^{l}_{mn}, $l = 0, 1/2, 1, 3/2,\ldots, -l \leqslant m, n \leqslant l$, form a basis of A. The structure of a C^* algebra is introduced into A in Woronowicz's approach[5,6] to the quantum group $\mathrm{SU}_q(2)$.

IX. MATRIX ELEMENTS OF THE REPRESENTATIONS T^l

Let us derive expressions for matrix elements of the irreducible reprsentations T^l of $U_q(\mathrm{su}_2)$ in terms of t_{11}, t_{12}, t_{21}, t_{22}. We put $l_2 = 1/2$ into (67b). It follows from (66a) that

$$\begin{bmatrix} l & 1/2 & l+1/2 \\ l & 1/2 & l+1/2 \end{bmatrix} = \begin{bmatrix} l & 1/2 & l+1/2 \\ -l & -1/2 & -l-1/2 \end{bmatrix} = 1.$$

Therefore,

$$d^{l+1/2}_{l+1/2,l+1/2} = d^{l}_{ll}t_{11},$$
$$d^{l+1/2}_{-l-1/2,l+1/2} = d^{l}_{-l,l}t_{21},$$
$$d^{l+1/2}_{l+1/2,-l-1/2} = d^{l}_{l,-l}t_{12},$$
$$d^{l+1/2}_{-l-1/2,-l-1/2} = d^{l}_{-l,-l}t_{22}.$$

We obtain from here that

$$d^{l}_{ll} = t^{2l}_{11}, \quad d^{l}_{-l,l} = t^{2l}_{21},$$
$$d^{l}_{l,-l} = t^{2l}_{12}, \quad d^{l}_{-l,-l} = t^{2l}_{22}. \tag{72}$$

Now let us put $l_1 = 1/2$ into (67a) and use expressions (66a) and (66b) for CGC's of the tensor product $T^{1/2} \otimes T^{1/2}$. We receive the relations

$$[2l+1]t_{11}d^{l}_{kj} = q^{-(2l-k-j)/4}([l+k+1][l+j+1])^{1/2} \times d^{l+1/2}_{k+1/2,j+1/2} + q^{(2l+k+j+2)/4} \times ([l-k][l-j])^{1/2}d^{l-1/2}_{k+1/2,j+1/2}, \tag{73}$$

$$[2l+1]t_{12}d^{l}_{kj} = q^{(k+j)/4}([l+k+1][l-j+1])^{1/2} \times d^{l+1/2}_{k+1/2,j-1/2} - q^{(k+j)/4} \times ([l-k][l+j])^{1/2}d^{l-1/2}_{k+1/2,jk-1/2}, \tag{74}$$

$$[2l+1]t_{21}d^{l}_{kj} = q^{(k+j)/4}([l-k+1][l+j+1])^{1/2} \times d^{l+1/2}_{k-1/2,j+1/2} - q^{(k+j)/4} \times ([l+k][l-j])^{1/2}d^{l-1/2}_{k-1/2,j+1/2}, \tag{75}$$

$$[2l+1]t_{22}d^{l}_{kj} = q^{(2l+k+j)/4}([l-k+1][l-j+1])^{1/2} \times d^{l+1/2}_{k-1/2,j-1/2} + q^{-(2l-k-j+2)/4} \times ([l+k][l+j])^{1/2}d^{l-1/2}_{k-1/2,j-1/2}. \tag{76}$$

In fact, these relations coincide with relations (4.6)–(4.9) from Ref. 17. Setting $l_2 = 1/2$ into (67a) we have

$$[2l+1]d^{l}_{kj}t_{11} = q^{(2l-k-j)/4}([l+k+1][l+j+1])^{1/2} \times d^{l+1/2}_{k+1/2,j+1/2} + q^{-(2l+k+j+2)/4} \times ([l-k][l-j])^{1/2}d^{l-1/2}_{k+1/2,j+1/2}, \tag{77}$$

$$[2l+1]d^{l}_{kj}t_{12} = q^{-(k+j)/4}([l+k+1][l-j+1])^{1/2} \times d^{l+1/2}_{k+1/2,j-1/2} - q^{-(k+j)/4} \times ([l-k][l+j])^{1/2}d^{l-1/2}_{k+1/2,j-1/2}, \tag{78}$$

$$[2l+1]d^{l}_{kj}t_{21} = q^{-(k+j)/4}([l-k+1][l+j+1])^{1/2} \times d^{l+1/2}_{k-1/2,j+1/2} - q^{-(k+j)/4} \times ([l+k][l-j])^{1/2}d^{l-1/2}_{k-1/2,j+1/2}, \tag{79}$$

$$[2l+1]d^{l}_{kj}t_{22} = q^{-(2l+k+j)/4}([l-k+1][l-j+1])^{1/2} \times d^{l+1/2}_{k-1/2,j-1/2} + q^{(2l-k-j+2)/4} \times ([l+k][l+j])^{1/2}d^{l-1/2}_{k-1/2,j-1/2}. \tag{80}$$

For $k = l$ relations (73) and (74) take the form

$$[2l+1]^{1/2}t_{11}d^{l}_{lj} = q^{-(l-j)/4}[l+j+1]^{1/2}d^{l+1/2}_{l+1/2,j+1/2}, \tag{81a}$$

$$[2l+1]^{1/2}t_{12}d^{l}_{lj} = q^{(l+j)/4}[l-j+1]^{1/2}d^{l+1/2}_{l+1/2,j-1/2}. \tag{81b}$$

If $k = -l$ in (85) and (86), then we have

$$[2l+1]^{1/2}t_{21}d^{l}_{-l,j} = q^{-(l-j)/4}[l+j+1]^{1/2}d^{l+1/2}_{-l-1/2,j+1/2}, \tag{82a}$$

$$[2l+1]^{1/2}t_{22}d^{l}_{-l,j} = q^{(l+j)/4}[l-j+1]^{1/2}d^{l+1/2}_{-l-1/2,j-1/2}. \tag{82b}$$

We consider relations (81a)–(82b) and similar relations, obtained from (77)–(80), as recurrence relations for matrix

elements $d^{l}_{\pm l,j}$ and $d^{l}_{j,\pm l}$. These recurrence relations and matrix elements (72) show that

$$d^{l}_{n,l} = q^{(l^2-n^2)/4}\left(\frac{[2l]!}{[l+n]![l-n]!}\right)^{1/2} t_{11}^{l+n} t_{21}^{l-n}, \tag{83}$$

$$d^{l}_{l,n} = q^{(l^2-n^2)/4}\left(\frac{[2l]!}{[l+n]![l-n]!}\right)^{1/2} t_{11}^{l+n} t_{12}^{l-n}, \tag{84}$$

$$d^{l}_{n,-l} = q^{(l^2-n^2)/4}\left(\frac{[2l]!}{[l+n]![1-n]!}\right)^{1/2} t_{12}^{l+n} t_{22}^{l-n}, \tag{85}$$

$$d^{l}_{-l,n} = q^{(l^2-n^2)/4}\left(\frac{[2l]!}{[l+n]![l-n]!}\right)^{1/2} t_{21}^{l+n} t_{22}^{l-n}. \tag{86}$$

Now expressions for general matrix elements of the representations T^l can be found. They have the form

$$\begin{aligned} d^{l}_{mn} = &\left(\frac{[l-n]![l+m]!}{[l-m]![l+n]!}\right)^{1/2} \\ &\times \frac{t_{11}^{m+n} t_{12}^{m-n}}{[m-n]!}\, q^{(m+n)(m-n)/4} \\ &\times {}_2\Phi_1(-l+m, l+m+1; m-n+1; q, \\ &-\sqrt{q} t_{12} t_{21}), \end{aligned} \tag{87}$$

$$\begin{aligned} d^{l}_{nm} = &\left(\frac{[l-n]![l+m]!}{[l-m]![l+n]!}\right)^{1/2} \\ &\times \frac{t_{11}^{m+n} t_{21}^{m-n}}{[m-n]!}\, q^{(m+n)(m-n)/4} \\ &\times {}_2\Phi_1(-l+m, l+m+1; m-n+1; q, \\ &-\sqrt{q} t_{12} t_{21}), \end{aligned} \tag{88}$$

if $m \geqslant |n|$, and the form

$$\begin{aligned} d^{l}_{-m,n} = &\left(\frac{[l+n]![l+m]!}{[l-m]![l-n]!}\right)^{1/2} \frac{q^{-(m+n)(m-n)/4}}{[m+n]!} \\ &\times {}_2\Phi_1(-l+m, l+m+1; m+n+1; q, \\ &-\sqrt{q} t_{12} t_{21}) t_{22}^{m-n} t_{21}^{m+n}, \end{aligned} \tag{89}$$

$$\begin{aligned} d^{l}_{n,-m} = &\left(\frac{[l+n]![l+m]!}{[l-m]![l-n]!}\right)^{1/2} \frac{q^{-(m+n)(m-n)/4}}{[m+n]!} \\ &\times {}_2\Phi_1(-l+m, l+m+1; m+n+1; q, \\ &-\sqrt{q} t_{12} t_{21}) t_{22}^{m-n} t_{12}^{m+n}, \end{aligned} \tag{90}$$

if $m \geqslant |n|$. Really, it follows from (83)–(86) that formulas (87)–(90) are valid for $m = l$. In order to prove their validity for general m we use a mathematical induction. Namely, we substitute expression (87) for d^{l}_{mn} into (73)–(75). After simple manipulations they reduce to formulas (28)–(30). But formulas (74) and (75) allow us to transform d^{l}_{mn} into d^{l+1}_{mn}. Therefore, expression (87) for d^{l}_{mn} is valid. Analogously, with the help of relations (73)–(75) we prove formula (88). Formulas (89) and (90) are proved by means of relations (78)–(80).

Other expressions for the matrix elements d^{l}_{mn} can be obtained from (87)–(90) with the help of relations (4), (5), and (8).

Remark: After the manuscript was submitted for publication, we learned of Biedenharn's papers[34,35] where a realization of the quantum algebra $U_q(\mathrm{su}_2)$ is constructed by means of a q-analog of the Jordan–Schwinger mapping. Using this realization a theory of q-tensor operators is developed. As in the classical case, this theory is a powerful tool for development of the theory of the Racah and Clebsch–Gordan coefficients. We also learned of Nomura's paper[36] where properties of the Clebsch–Gordan coefficients of $U_q(\mathrm{su}_2)$ are considered. In fact, there is no overlapping of our results with this paper.

[1] V. G. Drinfeld, Sov. Math. Dokl. **32**, 254 (1985).
[2] P. P. Kulish, N. Yu. Reshetikhin, and E. K. Sklyanin, Lett. Math. Phys. **5**, 395 (1981).
[3] E. K. Sklyanin, Funct. Anal. Appl. **16**, 27 (1982); **17**, 34 (1983).
[4] M. Jimbo, Lett. Math. Phys. **10**, 63 (1985); **11**, 247 (1986).
[5] S. L. Woronowicz, Commun. Math. Phys. **111**, 613 (1987).
[6] S. L. Woronowicz, Publ. RIMS, Kyoto Univ. **23**, 117 (1987).
[7] G. Lusztig, "Modular representations of quantum groups," MIT preprint, 1988.
[8] G. Lusztig, Adv. Math. **70**, 237 (1988).
[9] M. Rosso, Commun. Math. Phys. **117**, 581 (1988).
[10] M. Jimbo, "Field theory, quantum groups and strings," Lecture Notes in Physics, Vol. 246 (Springer-Verlag, Berlin, 1986), p. 334.
[11] A. J. Macfarlane, "On q-analogues of the quantum harmonic oscillator and the quantum group $SU(2)_q$", Cambridge Univ., England, preprint, 1989.
[12] L. Alvarez-Gaume, C. Gomez, and G. Sierra, "Duality and quantum groups," CERN preprint TH. 5369/89, 1989.
[13] H. J. de Vega, "Int. J. Mod. Phys. **4**, 2371 (1989).
[14] G. Moore and N. Seiberg, Commun. Math. Phys. **123**, 177 (1989).
[15] V. Pasquier and H. Saleur, "Symmetries of the XXZ chain and quantum groups," preprint SPht/88-187, Saclay, 1988.
[16] T. H. Koornwinder, Nederl. Acad. Wetensch. Proc. Ser. A **92**, 97 (1989).
[17] L. L. Vaksman and Ya. S. Soibelman, Funktsional Anal. Prilozhen. **22**(3), 1 (1988).
[18] A. N. Kirillov and N. Ju. Reshetikhin, "Representations of the algebra $U_q(\mathrm{su}(2))$, q-orthogonal polynomials and invariants of links," LOMI preprint E-9-88, Leningrad, 1988.
[19] H. T. Koelink and T. H. Koornwinder, Univ. of Leiden preprint, W-88-12, 1988.
[20] L. L. Vaksman, Dokl. Akad. Nauk, **306**, 269 (1989).
[21] L. C. Biedenharn, and J. D. Louck, *Angular Momentum in Quantum Physics* (Addison–Wesley, Reading, MA, 1981).
[22] T. Masuda, K. Mimachi, Y. Nakagami, M. Noimi, and K. Ueno, C. R. Acad. Sci. Paris, Ser. Math. **307**, 559 (1988).
[23] W. Hahn, Math. Nachr. **2**, 4, 263, 340 (1949), **3**, 257 (1950).
[24] H. Exton, *q-Hypergeometric Functions and Applications* (Ellis Hoorwood, Chichester, England, 1983).
[25] G. Gasper and M. Rahman, *Basic Hypergeometric Functions* (Cambridge, U.P. Cambridge, England, 1990).
[26] G. E. Andrews and R. Askey, in *Higher Combinatorics* (edited by M. Aigner (Reidel Dordrecht, 1977).
[27] E. G. Kalnins and W. Miller, Jr., SIAM J. Math. Anal. **19**, 1216 (1988).
[28] L. J. Slater, *Generalized Hypergeometric Functions* (Cambridge U. P., Cambridge, England, 1966).
[29] A. K. Agarwal, E. G. Kalnins, and W. Miller, Jr., SIAM J. Math. Anal. **18**, 1519 (1987).
[30] A. U. Klimyk and V. A. Groza, "Representations of quantum pseudounitary groups," preprint ITP-89-37P, Kiev, 1989.
[31] G. Andrews and R. Askey, "Classical orthogonal polynomials," Lecture Notes in Mathematics, Vol. 1171 (Springer-Verlag, Berlin, 1985), p. 36.
[32] R. Askey and J. Wilson, SIAM J. Math. Anal. **10**, 1008 (1979).
[33] V. Pasquier, Commun. Math. Phys. **118**, 355 (1988).
[34] L. C. Biedenharn, J. Phys. A: Math. Gen. **22**, L873 (1989).
[35] L. C. Biedenharn, "A q-boson realization of the quantum group $SU_q(2)$ and the theory of q-tensor operators," preprint, Duke Univ., Durham, NC, 1989.
[36] M. Nomura, J. Math. Phys. **30**, 2397 (1989).

Vector coherent state constructions of U(3) symmetric tensors and their SU(3)⊃SU(2)×U(1) Wigner coefficients

K. T. Hecht
Physics Department, University of Michigan, Ann Arbor, Michigan 48109

L. C. Biedenharn
Department of Physics, Duke University, Durham, North Carolina 27706

(Received 3 April 1990; accepted for publication 25 July 1990)

Generalized vector coherent state constructions of totally symmetric U(3) tensors are used to gain new expressions for the SU(3)⊃SU(2)×U(1) Wigner coefficients for the coupling $(\lambda_1\mu_1)\times(\lambda_2 0)\to(\lambda_3\mu_3)$. These expressions show how the extremely simple formulas of Le Blanc and Biedenharn, involving a single 9-*j* coefficient, arise as special cases of a general result that involves 12-*j* coefficients. A simpler general result involving only 9-*j* coefficients and *K*-normalization factors is derived in a way that can, in principle, be generalized to the generic coupling with multiplicity.

I. INTRODUCTION

In the past few years, a vector coherent state theory[1-5] (VCS) and its associated *K*-matrix technique[1,6,7] have been used to great advantage to evaluate explicit expressions for the matrix representations of higher rank Lie algebras. The unitary groups in the canonical chain $U(n)\supset U(n-1)\supset\cdots\supset U(2)\supset U(1)$ form a particularly simple example,[8] and VCS techniques have been used to cast many results for the U(*n*) Wigner–Racah calculus into new forms that reveal the structure of the Wigner and recoupling coefficients in a new light. It has been shown in particular[9] that the U(*n*) elementary unit projective operators of Biedenharn and Louck[10] can be written down very simply in terms of U($n-1$) Racah coefficients and the simple *K*-normalization factors of VCS theory. Very recently Le Blanc and Biedenharn[11] have shown that some classes of $U(n)\supset U(n-1)\times U(1)$ reduced Wigner coefficients are simply products of U($n-1$)9-*j* type recoupling coefficients and *K*-normalization factors. The question naturally arises: To what extent can the most general $U(n)\supset U(n-1)\times U(1)$ reduced Wigner coefficient be expressed in terms of U($n-1$) recoupling coefficients and the *K*-normalization factors of VCS theory? The earliest detailed applications of VCS theory have focused on the matrix elements of the generators of the algebra. Very recently[12,13] VCS theory has been generalized to include the Bargmann space realizations of more general operators lying outside the algebra. This generalization now makes it possible to examine the spectacularly simple class of U(*n*) tensors of Le Blanc and Biedenharn[11] and show that they are a special case of a more general result. In this generalization, $U(n)\supset U(n-1)\times U(1)$ reduced Wigner coefficients are expressible in terms of summations involving U($n-1$) 12-*j* type recoupling coefficients. For the Le Blanc–Biedenharn case, these sums collapse to a single term in which the 12-*j* type coefficient collapses to a 9-*j* type coefficient. The ultimate aim of constructing a U(*n*) tensor operator calculus in a unique (author-independent) way has not been fully implemented in the generic case with multiplicity.[14] It may therefore be useful to first reexamine the special multiplicity-free case of totally symmetric U(*n*) tensors within the framework of the generalized VCS theory.[12,13] It is the purpose of this investigation to generalize the Le Blanc–Biedenharn result. To avoid the multiplicity problem, however, the investigation is restricted to the special case of totally symmetric tensors. To avoid some of the notational complexities of the Gel'fand notation needed for general U(*n*), a further simplification to $n=3$ is made so that the U($n-1$) recoupling coefficients are expressible in terms of well-known angular momentum recoupling coefficients of 12-*j*, 9-*j*, or 6-*j* type. Three new expressions are given for the SU(3)⊃SU(2)×U(1) reduced Wigner coefficients for the most general coupling of type $(\lambda_1\mu_1)\times(\lambda_2 0)\to(\lambda_3\mu_3)$. In terms of their complexity and the number of required summations, these expressions are comparable to previously known[15] results. Since all results are expressed in terms of SU(2) recoupling coefficients and the *K*-normalization factors of VCS theory, these results reveal the structure of the SU(3)⊃SU(2)×U(1) reduced Wigner coefficients in a new light. They are derived by VCS techniques that can, in principle, be generalized to the generic case with multiplicity, the ultimate aim of this type of investigation.

The paper is organized in the following way. Section II gives the Bargmann space realization of totally symmetric U(3) tensor operators using the generalized VCS approach.[12] In this approach, a U(3) tensor operator is factored into two parts in an SU(2)-coupled basis: (1) an "intrinsic" operator acting only on the generalized VCS "vacuum" states, U(1) extremal states in the U(3)⊃SU(2)×U(1) scheme; and (2) a Bargmann space (**z**-space) operator that changes the U(1) weights. As soon as the "intrinsic" operator reduced matrix elements are known, the evaluation of SU(3)⊃SU(2)×U(1) reduced Wigner coefficients is reduced to an exercise in angular momentum coupling. The "intrinsic" operator reduced matrix elements are evaluated in Sec. III. The new expression for the SU(3)⊃SU(2)×U(1) reduced Wigner coefficient is then

 0022-2488/90/122781-16$03.00

given in Sec. IV. For the coupling $(\lambda_1\mu_1)\times(\lambda_2 0)\to(\lambda_3\mu_2)$. The expression of Sec. IV is particularly simple in practice if n, the number of squares added to row 3 of the Young tableau for $(\lambda_1\mu_1)$ is very small compared with $(\lambda_2 - n)$ the number of squares added to rows 1 and 2. The case $n=0$ leads to the first Le Blanc–Biedenharn result. For the case when $(\lambda_2 - n)$ is small an alternate but similar expression may be more sueful. This is given in Sec. V. For the case $(\lambda_2 - n)=0$ it collapses to the second Le Blanc–Biedenharn result. The general expressions of both Sec. IV and V involve 12-j coefficients. An even simpler expression, involving only 9-j coefficients, is derived in Sec. VI by a buildup process that compounds the two special Le Blanc–Biedenharn results. This final approach not only gives the simplest result from the point of view of actual computations but also shows the greatest promise for the needed generalization to the generic case with multiplicity.

II. VCS REALIZATION OF TOTALLY SYMMETRIC U(3) TENSORS

The U(3) generators E_{ij} can be realized in the usual way in terms of oscillator creation and annihilation operators, $\alpha_{ip}^\dagger$ and α_{ip}; with "spatial" index $i=1, 2, 3$, or x, y, z, and "particle index p, with $p=1, \ldots, n$:

$$E_{ij}=\frac{1}{2}\sum_{p=1}^{n}(\alpha_{ip}^\dagger\alpha_{jp}+\alpha_{jp}\alpha_{ip}^\dagger). \tag{1}$$

The complementary[16] U(n) generators C_{pq} are

$$C_{pq}=\frac{1}{2}\sum_{i=1}^{3}(\alpha_{ip}^\dagger\alpha_{iq}+\alpha_{iq}\alpha_{ip}^\dagger). \tag{2}$$

For U(3) it is sufficient to choose $n=3$, and this choice will be made. However, the specific value of n plays very little role in the present investigation. [A restriction to SU(3) with $n=2$ has been shown to have some advantages by Le Blanc and Rowe[17] but would require some modification in the present construction.]

In the VCS theory, the U(3) generators are organized into (1) an Abelian nilpotent algebra of raising operators $E_{i3}\equiv A_i$, with $i=1, 2$; (2) an Abelian nilpotent algebra of lowering operators $E_{3i}=A_i^\dagger$, with $i=1, 2$; (3) The U(2) subgroup generators E_{ij} with $i,j=1, 2$; (4) the U(1) subgroup generator E_{33}.

The generators of U(2) $\otimes$ U(1) are called the core subgroup generators.

The U(3) state vectors can be specified by the Young frame integers $[m_{13}\ m_{23}\ m_{33}]$ with standard Gel'fand subgroup labels m_{12}, m_{22}, and m_{11}. Alternatively, they can be specified by the total number of oscillator quanta $N=m_{13}+m_{23}+m_{33}$, the Cartan SU(3) labels $\lambda=m_{13}-m_{23}$, $\mu=m_{23}-m_{33}$, and U(2)$\times$U(1) subgroup labels given in the notation of Ref. 5 by angular momentum quantum numbers of I, M_I, and the U(1) label Y, the eigenvalue of $\frac{1}{3}(E_{11}+E_{22}-2E_{33})$, with $Y=\frac{1}{3}(\lambda+2\mu)-w$, $w=0, 1, \ldots, \lambda+\mu$. Note that $w=m_{13}+m_{23}-m_{12}-m_{22}$, $I=\frac{1}{2}(m_{12}-m_{22})$, $M_I=m_{11}-\frac{1}{2}(m_{12}+m_{22})$. Note also that w gives the eigenvalue of E_{33}. The set of generalized "vacuum" or "intrinsic" states of VCS theory will be chosen to be the states with $w=0$, $|N(\lambda\mu)w=0, I=\lambda/2, M_I=m\rangle$, for which

$$E_{i3}|N(\lambda\mu)w=0,\quad I=(\lambda/2)m\rangle=0,$$
$$\text{for } i=1,2;\ m=+\lambda/2,\ldots,-\lambda/2. \tag{3}$$

Note that the operators E_{i3} are *raising* operators for Y. The state with $w=0$, $m=\frac{1}{2}\lambda$ is a *highest* weight state. In terms of the Elliott label[18] $\epsilon=-3Y$, however, the E_{i3} become *lowering* operators and the vacuum states become *lowest* ϵ-weight states. The words raising and lowering will therefore be avoided. The $E_{i3}\equiv A_i$ will be named annihilation operators instead since they annihilate the generalized vacuum states of Eq. (3), whereas the $E_{3i}\equiv A_i^\dagger$ can be named creation operators. In the U(3)$\times$U(n) realization, with $n=3$, the generalized vacuum state with $m=+\frac{1}{2}\lambda$ has the form

$$|N(\lambda\mu)w=0,\quad I=m=\lambda/2\rangle$$
$$=\mathcal{N}_H(\alpha_{11}^\dagger)^\lambda\begin{vmatrix}\alpha_{11}^\dagger & \alpha_{12}^\dagger\\ \alpha_{21}^\dagger & \alpha_{21}^\dagger\end{vmatrix}^\mu \times\begin{vmatrix}\alpha_{11}^\dagger & \alpha_{12}^\dagger & \alpha_{13}^\dagger\\ \alpha_{21}^\dagger & \alpha_{22}^\dagger & \alpha_{23}^\dagger\\ \alpha_{31}^\dagger & \alpha_{32}^\dagger & \alpha_{33}^\dagger\end{vmatrix}^{m_{33}}, \tag{4a}$$

with

$$\mathcal{N}_H=\sqrt{\frac{(\lambda+1)(\mu+1)(\lambda+\mu+2)}{(\lambda+\mu+m_{33}+2)!(\mu+m_{33}+1)!m_{33}!}}. \tag{4b}$$

The vector coherent state is built in terms of two complex variables z_i $(i=1,2)$, through the action of the creation operators E_{3i} on the generalized vacuum or intrinsic states:

$$|\mathbf{z};N(\lambda\mu)m\rangle=\exp(z_1^*E_{31}+z_2^*E_{32})|N(\lambda\mu)w=0,(\lambda/2)m\rangle. \tag{5}$$

Note that this coherent state carries the labels $N(\lambda\mu)$ and $m=+\lambda/2,\ldots,-\lambda/2$. It forms a $(\lambda+1)$-dimensional array, i.e., it is a vector quantity.

State vectors can be expanded in terms of the U(3) basis vectors $|N(\lambda\mu)wIM_I\rangle$. In the VCS method these are mapped into their z-space functional realizations:

$$|N(\lambda\mu)wIM_I\rangle\to|N(\lambda\mu)wIM_I\rangle_{\rm VCS},$$
$$|N(\lambda\mu)wIM_I\rangle_{\rm VCS}$$
$$=\sum_m\langle N(\lambda\mu)0(\lambda/2)m|e^{\mathbf{z}\cdot\mathbf{A}}|N(\lambda\mu)wIM_I\rangle \times|N(\lambda\mu)0(\lambda/2)m\rangle\otimes|0\rangle, \tag{6}$$

with $\mathbf{z}\cdot\mathbf{A}=z_iA_i=z_1E_{13}+z_2E_{23}$. Note that this is a linear combination of intrinsic space standard kets, $|N(\lambda\mu)0(\lambda/2)m\rangle$, with coefficients that are functions of $\mathbf{z}$. For completeness the z-space vacuum ket $|0\rangle$ (for the action of the z bosons), has been included in Eq. (6). However, since its z-space realization is the simple number 1 it will usually be omitted throughout later sections.

Operators $\mathbf{O}$ are mapped into their z-space realizations $\Gamma(\mathbf{O})$ via

$$(\mathbf{O}|N(\lambda\mu)wIM_I\rangle)_{\rm VCS} + \sum_m \langle N(\lambda\mu)0\,\frac{\lambda}{2}m|e^{z\cdot\mathbf{A}}\mathbf{O}e^{-z\cdot\mathbf{A}}e^{z\cdot\mathbf{A}}|N(\lambda\mu)wIM_I\rangle|N(\lambda\mu)0\,\frac{\lambda}{2}m\rangle\otimes|0\rangle$$

$$= \sum_m \sum_{\bar{w}\bar{I}\bar{M}_I} \langle N(\lambda\mu)0(\lambda/2)m|\Gamma(\mathbf{O})|N(\lambda\mu)\bar{w}\bar{I}\bar{M}_i\rangle\langle N(\lambda\mu)\bar{w}\bar{I}\bar{M}_I|e^{z\cdot\mathbf{A}}|N(\lambda\mu)wIM_I\rangle|N(\lambda\mu)0\frac{\lambda}{2}m\rangle\otimes|0\rangle, \tag{7}$$

with

$$\Gamma(\mathbf{O}) = \mathbf{O} + [\mathbf{z}\cdot\mathbf{A},\mathbf{O}] + \tfrac{1}{2}[\mathbf{z}\cdot\mathbf{A},[\mathbf{z}\cdot\mathbf{A},\mathbf{O}]] + \cdots. \tag{8}$$

The z-space realizations for the generators E_{ij} are given in Refs. 5 and 8, but are repeated here for completeness.

With $i,j = 1,2$,

$$\Gamma(E_{i3}) = \Gamma(A_i) = \frac{\partial}{\partial z_i},$$

$$\Gamma(E_{ij}) = \mathbb{E}_{ij} - z_j\frac{\partial}{\partial z_i} = E_{ij}^{\rm intr} + E_{ij}^{\rm coll.} \tag{9}$$

$$\Gamma(E_{3i}) = \Gamma(A_i^\dagger) = \sum_{\alpha=1}^{2}\left(\mathbb{E}_{\alpha i}z_\alpha - \mathbb{E}_{33}z_i - z_iz_\alpha\frac{\partial}{\partial z_\alpha}\right).$$

Note that they are functions of the z-space operators z_i, $\partial/\partial z_j$, and intrinsic operators $\mathbb{E}_{ij}$ (denoted by double lines). These intrinsic operators are defined only through their action on the intrinsic states. They commute with the z-space operators z_i, $\partial/\partial z_j$, and in the matrix element of $\Gamma(\mathbf{O})$ of Eq. (7) they must be worked through to the left so that they can act on the intrinsic state. Since generators do not change the irreducible representation of the group, the $\mathbb{E}_{ij}$ connect intrinsic states only to intrinsic states (possibly with $m' \neq m$). The $\mathbb{E}_{ij}$ are thus defined through their pure intrinsic state matrix elements, e.g., with $\mathbb{E}_{12} \equiv \mathbb{I}_+$,

$$\langle N(\lambda\mu)0(\lambda/2)m|\mathbb{E}_{12}|N(\lambda\mu)0(\lambda/2)(m-1)\rangle$$
$$= \sqrt{(\lambda/2+m)(\lambda/2-m+1)}. \tag{10}$$

In the generalized VCS method[12,13] operators $\mathbf{O}$ outside the group algebra are to be included. Since the action of such operators can change the U(3) irreducible representations, Eq. (7) is to be generalized to

$$(\mathbf{O}|N(\lambda\mu)wIM_I\rangle)_{\rm VCS}$$
$$= \sum_{N'(\lambda'\mu')}\sum_{m'}\sum_{\bar{w}\bar{I}\bar{M}_I}\langle N'(\lambda'\mu')0\,\frac{\lambda'}{2}m'|$$
$$\times\Gamma(\mathbf{O})|N(\lambda\mu)\bar{w}\bar{I}\bar{M}_I\rangle$$
$$\times\langle N(\lambda\mu)\bar{w}\bar{I}\bar{M}_I|e^{z\cdot\mathbf{A}}|N(\lambda\mu)wIM_I\rangle$$
$$\times|N'(\lambda'\mu')0(\lambda'/2)m'\rangle\otimes|0\rangle. \tag{11}$$

The oscillator creation and annihilation operators $(\alpha_{ia}^\dagger, \alpha_{ia}$, for specific particle index a are of particular interest since more complicated operators can be built from these. We note first that the operators $\alpha_{ia}^\dagger$, $E_{i3} = A_i, z_i^*$, $(i = 1,2)$, transform as U(2) tensors of rank [10], whereas their conjugate partners α_{ia}, $E_{3i} = A_i^\dagger$, z_i transform as U(2) tensors [0–1]. In terms of standard spherical tensor T^I_M, therefore,

$$\{z_1^*, z_2^*\} = \{Z^{1/2}_{+1/2}(\mathbf{z}^*), Z^{1/2}_{-1/2}(\mathbf{z}^*)\}, \tag{12a}$$

whereas

$$\{z_1,z_2\} = \{+Z^{1/2}_{-1/2}(\mathbf{z}), -Z^{1/2}_{+1/2}(\mathbf{z})\}$$
$$= \{(-1)^{1/2-m}Z^{1/2}_{-m}(\mathbf{z})\}, \tag{12b}$$

and

$$\{\alpha_{1a}^\dagger,\alpha_{2a}^\dagger\} = \{(\alpha^\dagger)^{1/2}_{+1/2},(\alpha^\dagger)^{1/2}_{-1/2}\};$$
$$\{\alpha_{1a},\alpha_{2a}\} = \{(\tilde{\alpha})^{1/2}_{-1/2}, -(\tilde{\alpha}^{1/2}_{+1/2}\}. \tag{13}$$

Note also that $\alpha_{3a}^\dagger$ and α_{3a} are SU(2) scalars.

The z-space realizations of the operators $\alpha_{ia}^\dagger$, α_{ia} can now be constructed by the application of Eq. (8) to yield

$$\Gamma((\alpha_a^\dagger)^{1/2}_m) = (\mathcal{A}_a^\dagger)^{1/2}_m, \tag{14a}$$

$$\Gamma(\alpha_{3a}^\dagger) = \mathcal{A}_{3a}^\dagger + \mathcal{A}_{1a}^\dagger z_1 + \mathcal{A}_{2a}^\dagger z_2$$
$$= (\mathcal{A}_a^\dagger)^0_0 - \sqrt{2}[\mathcal{A}_a^{\dagger 1/2}\times\mathbb{Z}^{1/2}(\mathbf{z})]^0_0, \tag{14b}$$

$$\Gamma(\alpha_{3a}) = \mathcal{A}_{3a} = (\mathcal{A}_a)^0_0, \tag{15a}$$

$$\Gamma((\alpha_a)^{1/2}_m) = (\mathcal{A}_a)^{1/2}_m - Z^{1/2}_m(\mathbf{z})(\mathcal{A}_a)^0_0. \tag{15b}$$

After the application of Eq. (8) the operators $\alpha_{ia}^\dagger$, α_{ia} have been formally replaced by double-line intrinsic operators. These double-line operators again commute with the z_i and $\partial/\partial z_j$. They are again to be worked through to the left in the matrix element of $\Gamma(\mathbf{O})$ in Eq. (11) where they are then defined through their actions on the adjacent intrinsic state. Unlike the matrix elements of Eq. (10), however, they can convert an intrinsic state with $N'(\lambda'\mu')$ to a nonintrinsic state, (with $\bar{w}\neq 0$), in the representation with $N(\lambda\mu)$. The practical application of VCS theory thus depends on the evaluation of the matrix elements of double-line operators such as $\mathcal{A}^\dagger$, $\mathcal{A}$ between the purely intrinsic states on the left and the permitted states on the right (see Sec. III). Finally, note that the square bracket in Eq. (14) denotes angular momentum coupling using a *right to left* coupling order. This right to left coupling order convention simplifies phase factors in the VCS constructions and will be used throughout.

The VCS mappings of Eqs. (6) and (7) are nonunitary. The $\Gamma(\mathbf{O})$ are, in general, nonunitary realizations of the operator $\mathbf{O}$. Clearly, $\Gamma(A_i^\dagger) \neq (\Gamma(A_i))^\dagger$; and $\Gamma(\alpha_{i\alpha}^\dagger) \neq (\Gamma(\alpha_{ia}))^\dagger$. Similarly, the VCS state vector $|N(\lambda\mu)wIM_I\rangle_{\rm VCS}$ of Eq. (6) is not normalized. It will be instructive to give a specific evaluation of $|N(\lambda\mu)wIM_I\rangle_{\rm VCS}$, through

$$\langle N(\lambda\mu)0(\lambda/2)m|e^{z\cdot\mathbf{A}}|N(\lambda\mu)wIM_I\rangle$$
$$= \langle N(\lambda\mu)wIM_I|e^{\mathbf{z}^*\cdot\mathbf{A}^\dagger}|N(\lambda\mu)0(\lambda/2)m\rangle^*$$
$$= \sum_{k,m_k}\langle N(\lambda\mu)wIM_I|Z^{k/2}_{m_k}(\mathbf{A}^\dagger)$$
$$\times|N(\lambda\mu)0(\lambda/2)m\rangle^*Z^{k/2}_{m_k}(\mathbf{z}), \tag{16}$$

where we have used

$$\mathbf{z}^*\cdot\mathbf{A}^\dagger = z_1^*E_{31} + z_2^*E_{32} = \sqrt{2}[\mathbf{A}^{\dagger 1/2}\times Z^{1/2}(\mathbf{z}^*)]^0_0, \tag{17}$$

and repeated use of the angular momentum recoupling transformation:

$$[[Z^{(k-1)/2}(\mathbf{A}^\dagger)\times Z^{(k-1)/2}(\mathbf{z}^*)]^0\times[\mathbf{A}^{\dagger 1/2}\times Z^{1/2}(\mathbf{z}^*)]^0]^0_0$$
$$=\begin{bmatrix} 1/2 & 1/2 & 0 \\ (k-1)/2 & (k-1)/2 & 0 \\ k/2 & k/2 & 0 \end{bmatrix}$$
$$\times[[Z^{(k-1)/2}(\mathbf{A}\dagger)\times Z^{1/2}(\mathbf{A}\dagger)]^{k/2}$$
$$\times[Z^{(k-1)/2}(\mathbf{z}^*)\times Z^{1/2}(\mathbf{z}^*)]^{k/2}]^0_0 \qquad (18)$$
$$=\sqrt{\frac{k(k+1)}{2}}\,[Z^{k/2}(\mathbf{A}\dagger)\times Z^{k/2}(\mathbf{z}^*)]^0_0.$$

In Eq. (18) we have used the value of the unitary (square bracket) form of the 9-j recoupling coefficient and the buildup relation for the z-space boson polynomials:

$$[Z^{w_1/2}(\mathbf{z})\times Z^{w_2/2}(\mathbf{z})]^{w/2}_m$$
$$=\delta_{w,w_1+w_2}\sqrt{(w_1+w_2)!/w_1!w_2!}\,Z^{w/2}_m(\mathbf{z}). \qquad (19)$$

In Eq. (16) we have also used the conjugation relation

$$\left(Z^{k/2}_{-m_k}(\mathbf{z}^*)\right)^* = (-1)^{k/2+m_k}Z^{k/2}_{m_k}(\mathbf{z}). \qquad (20)$$

The buildup relation together with Eq. (12) leads to the specific construction

$$Z^{w/2}_m(\mathbf{z}) = \frac{(z_1)^{w/2-m}(-z_2)^{w/2+m}}{\sqrt{(w/2-m)!(w/2+m)!}}, \qquad (21)$$

where this is an eigenfunction of $\langle I_0\rangle^{\text{coll}} = \frac{1}{2}(E^{\text{coll}}_{11} - E^{\text{coll}}_{22})$ and $(\mathbf{I}^{\text{coll}}\cdot\mathbf{I}^{\text{coll}})$ with eigenvalues m and $(w/2)(w/2+1)$; [see Eq. (9)]. The creation operator polynomial $Z(\mathbf{A}^\dagger)$ is obtained from $Z(\mathbf{z})$ by the replacement $z_i \to A^\dagger_i$. The *non*-normalized state $|\phi_{\text{wIM}}\rangle$, constructed through the action of w creation operators $\mathbf{A}^\dagger$, via

$$|\phi_{wIM_I}\rangle = [Z^{w/2}(\mathbf{A}^\dagger)\times|N(\lambda\mu)0(\lambda/2)\rangle]^I_{MI}$$
$$=\sum_m \langle(\lambda/2)m(w/2)(M_I-m)|IM_I\rangle$$
$$\times Z^{w/2}_{(M_I-m)}(\mathbf{A}^\dagger)|N(\lambda\mu)0(\lambda/2)m\rangle \qquad (22a)$$

is orthogonal to states $|N(\lambda\mu)w'I'M'_I\rangle$ with $w'\neq w$, or $I'\neq I$ $M'_I\neq M_I$ but is not normalized. The normalization factor is given by the K-matrix element of VCS theory[5,8] [for the derivation, see Eq. (28) below]:

$$|N(\lambda\mu)wIM_I\rangle = [1/K(\lambda\mu)_{wI}]\,|\phi_{wIM_I}\rangle. \qquad (22b)$$

Equations (6), (16), and (22) thus lead to

$$|N(\lambda\mu)wIM_I\rangle_{\text{VCS}}$$
$$=K(\lambda\mu)_{wI}\sum_m\langle\frac{\lambda}{2}\,m\frac{w}{2}(M_I-m)|IM_I\rangle$$
$$\times Z^{w/2}_{M_I-m}(\mathbf{z})|N(\lambda\mu)0(\lambda/2)m\rangle\otimes|0\rangle$$
$$=K(\lambda\mu)_{wI}|N(\lambda\mu)wIM_I). \qquad (23)$$

The normalized z-space state vector will henceforth be denoted by

$$|N(\lambda\mu)wIM) = \left[Z^{w/2}(\mathbf{z})\times|(\lambda\mu)0\frac{\lambda}{2}\right)\Big]^I_M$$
$$\equiv|(\lambda\mu)\left[\frac{\mathrm{w}}{2}\times\frac{\lambda}{2}\right]\mathrm{I,M}\Big). \qquad (24)$$

Note, in particular, that the state vector in Eq. (24) has been written with a round parenthesis. (Note also that the z-space vacuum vector $|0\rangle$ will be omitted henceforth for economy of notation; for the same reason the label N will henceforth be omitted but will be quietly understood.) The z-space state vectors of Eq. (24) form an orthonormal set with respect to the z-space integrations with the standard Bargmann measure.[5] In evaluating matrix elements, it will be very useful to indicate explicitly whether matrix elements are to be calculated through their z-space integrations (with the Bargmann measure) or in standard Hilbert space form. For this reason the state vector of Eq. (24) has been written with a round parenthesis, $|\cdots)$, whereas the state vectors of Eq. (22) are interpreted as standard Hilbert space vectors and are denoted by angular brackets, $|\cdots\rangle$. The appearance of round parentheses in a matrix element henceforth will automatically signal z-space integrations and pure intrinsic space operations. The appearance of angular brackets on the other hand, will signal standard Hilbert space operations. To transcribe the matrix element of an operator $\mathbf{O}$ between states of the orthonormal Hilbert space basis $|N(\lambda\mu)wIM_I\rangle$ to the corresponding z-space matrix element, we not only need to transcribe to the orthonormal z-space basis $|N(\lambda\mu)wIM_I)$ but also need to transform the nonunitary z-space realization $\Gamma(\mathbf{O})$ to a unitary realization of the operator to be denoted by $\gamma(\mathbf{O})$. In the VCS technique, this transformation is achieved via the K operator:[5]

$$\gamma(\mathbf{O}) = K^{-1}\Gamma(\mathbf{O})K = (\gamma(\mathbf{O}^\dagger))^\dagger. \qquad (25)$$

The Hilbert space matrix element of $\mathbf{O}$ can thus be transcribed to the z-space matrix element of $\gamma(\mathbf{O})$:

$$\langle(\lambda'\mu')w'I'M'|\mathbf{O}|(\lambda\mu)wIM\rangle$$
$$=((\lambda'\mu')w'I'M'|\gamma(\mathbf{O})|(\lambda\mu)wIM)$$
$$=K^{-1}(\lambda'\mu')_{w'I'}\Big((\lambda'\mu')\left[\frac{w'}{2}\times\frac{\lambda'}{2}\right]I'M'\Big|$$
$$\times\Gamma(\mathbf{O})\Big|(\lambda\mu)\left[\frac{w}{2}\times\frac{\lambda}{2}\right]IM\Big)K(\lambda\mu)_{wI}. \qquad (26)$$

This is a basic relation that will be used repeatedly to calculate matrix elements of intrinsic operators. It can also be used to verify the normalization factor character of K in Eq. (22). Note that

$$\langle\phi_{wIM}|(\lambda'\mu')w'I'M'\rangle$$
$$=\sum_m\left\langle\frac{\lambda}{2}\,(M-m)\,\frac{w}{2}\,m\Big|IM\right\rangle$$
$$\times(-1)^{w/2-m}\langle(\lambda\mu)0(\lambda/2)(M-m)|$$
$$\times Z^{w/2}_{-m}(\mathbf{A})|(\lambda'\mu')w'I'M'\rangle. \qquad (27a)$$

where we have used the conjugation properties of the Z of Eq. (20) to obtain $(Z^{w/2}_m(\mathbf{A}^\dagger))^\dagger = (-1)^{w/2-m}Z^{w/2}_{-m}(\mathbf{A})$. Equation (26) can then be used to give

$$\langle(\lambda\mu)0(\lambda/2)(M-m)|Z^{w/2}_{-m}(\mathbf{A})|(\lambda'\mu')w'I'M'\rangle$$
$$=((\lambda\mu)0(\lambda/2)(M-m)|$$
$$\times K^{-1}\Gamma(Z^{w/2}_{-m}(\mathbf{A}))K|(\lambda'\mu')w'I'M')$$
$$=1\times\Big((\lambda\mu)0\frac{\lambda}{2}\,(M-m)|\Big(Z^{w/2}_{-m}\left(\frac{\partial}{\partial\mathbf{z}}\right)\Big)$$
$$\times|(\lambda'\mu')w'I'M'\Big)K(\lambda'\mu')_{w'I'}$$

$$= (-1)^{w/2-m}((\lambda'\mu')w'I'M'|Z_m^{w/2}(\mathbf{z})|(\lambda\mu)w$$
$$= 0(\lambda/2)(M-m))^* K(\lambda'\mu')_{w'I'}, \tag{27b}$$

where we have used $\Gamma(A_i) = \partial/\partial z_i$, and the fact that $\partial/\partial z_i$ is the adjoint of z_i with respect to Bargmann integrations.[5] Finally, K or K^{-1} acts as a simple unit operator on the normalized intrinsic state. Equations (27a) and (27b), with the orthonormality of the states (24), then give

$$\langle\phi_{wIM}|(\lambda'\mu')w'I'M'\rangle$$
$$= ((\lambda'\mu')w'I'M'|(\lambda\mu)wIM)^* K(\lambda'\mu')_{w'I'}$$
$$= \delta_{(\lambda'\mu')(\lambda\mu)}\delta_{w'w}\delta_{I'I}\delta_{M'M}K(\lambda\mu)_{wI}, \tag{28a}$$

leading to

$$\langle\phi_{wIM}|\phi_{wIM}\rangle$$
$$= \sum_{(\lambda'\mu')}\sum_{w'I'M'}\langle\phi_{wIM}|(\lambda'\mu')w'I'M'\rangle$$
$$\times\langle(\lambda'\mu')w'I'M'|\phi_{wIM}\rangle$$
$$= |K(\lambda\mu)_{wI}|^2. \tag{28b}$$

The K-factors have been evaluated by VCS techniques:[5,8]

$$K(\lambda\mu)_{wI} = \sqrt{\frac{(\lambda+\mu+1)!\mu!}{(\lambda/2+\mu+1-w/2+I)!(\mu+\lambda/2-w/2-I)!}}, \tag{29a}$$

or, with $w = p+q$, $I = \frac{1}{2}\lambda - \frac{1}{2}p + \frac{1}{2}q$,

$$K(\lambda\mu)_{pq} = \sqrt{\frac{(\lambda+\mu+1)!\mu!}{(\lambda+\mu+1-p)!(\mu-q)!}}. \tag{29b}$$

General U(3) tensors can be constructed from the oscillator creation and annihilation operators so that their VCS realizations follow from Eqs. (14) and (15). Note, however, that Eqs. (14) and (15) lead to the following form for the VCS realization of the group generators $E_{i3} = A_i$ of annihilation type: $\Gamma(A_i) = \Sigma_a \phi_{ia}^\dagger \phi_{3a}$; i.e., they are built entirely from intrinsic (double-line) operators. This is quite different from the "standard" realization of these group generators,[5,8] $\Gamma(A_i) = \partial/\partial z_i$, in which they are built from pure z-space operators. It is well known that coherent state realizations of operators are not unique due to the overcompleteness of coherent states. Both realizations must, however, give the same matrix elements. This was demonstrated explicitly in Ref. 13 for the analogous versions of the annihilation generators of the Sp(6) $\supset$ U(3) algebra. It is to be noted, however, that the structure of the expressions for the matrix elements can be very simple for one type of coherent state realization and very cumbersome for another. In some cases a search for an optimal realization may therefore be needed.

Totally symmetric U(3) tensor operators can be constructed through the three-dimensional oscillator creation operators in a single particle variable a, say $a = 1$, via the polynomials

$$T^{(\lambda_2 0)}_{w_2 I_2 M_2}(\alpha_a^\dagger) = \left[(\alpha_{3a}^\dagger)^{w_2}/\sqrt{w_2!}\right] P^{I_2 = (1/2)(\lambda_2 - w_2)}_{M_2}(\alpha_a^\dagger), \tag{30a}$$

with

$$P^{I_2}_{M_2}(\alpha_a^\dagger) = \frac{(\alpha_{1a}^\dagger)^{I_2+M_2}}{\sqrt{(I_2+M_2)!}}\frac{(\alpha_{2a}^\dagger)^{I_2-M_2}}{\sqrt{(I_2-M_2)!}}. \tag{30b}$$

(The subscript 2 on the quantum numbers λ, w, I, M is used for later applications.) The VCS realizations of these totally symmetric U(3) tensors could be obtained by repeated application of Eqs. (14) or preferably by the direct application of Eq. (8).

For the latter, the needed tools are the commutator

$$[\mathbf{z}\cdot\mathbf{A}, \alpha_{3a}^\dagger] = \alpha_{1a}^\dagger z_1 + \alpha_{2a}^\dagger z_2 = -\sqrt{2}[\alpha^{\dagger 1/2}\times Z^{1/2}(\mathbf{z})]_0^0, \tag{31}$$

and the angular momentum recoupling transformation

$$[[P^{w/2}(\alpha^\dagger)\times Z^{w/2}(\mathbf{z})]^0\times[\alpha^{\dagger 1/2}\times Z^{1/2}]^0]_0^0$$
$$= \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{w}{2} & \frac{w}{2} & 0 \\ \frac{w+1}{2} & \frac{w+1}{2} & 0 \end{bmatrix}$$
$$\times[[P^{w/2}(\alpha^\dagger)\times(\alpha^\dagger)^{1/2}]^{(w+1)/2}$$
$$\times[Z^{w/2}(\mathbf{z})\times Z^{1/2}]^{(w+1)/2}]_0^0$$
$$= \sqrt{(w+2)(w+1)/2}[P^{(w+1)/2}(\alpha^\dagger)$$
$$\times Z^{(w+1)/2}(\mathbf{z})]_0^0, \tag{32}$$

where we have again used the value of the unitary (square bracket) form of the 9-j recoupling coefficient and the build-up relation (19), together with the analogous relation

$$[P^{w_1/2}(\alpha^\dagger)\times P^{w_2/2}(\alpha^\dagger)]_m^{w/2}$$
$$= \delta_{w,w_1+w_2}\sqrt{(w_1+w_2)!/w_1!w_2!}P_m^{w/2}(\alpha^\dagger). \tag{33}$$

Repeated applications of Eqs. (31) and (32) in the νth commutator of Eq. (8) yields

$$\Gamma(T^{(\lambda_2 0)}_{w_2 I_2 = (1/2)(\lambda_2 - w_2) M_2}(\alpha_a^\dagger))$$
$$= \Gamma\left(\frac{(\alpha_{3a}^\dagger)^{w_2}}{\sqrt{w_2!}}P^{(1/2)(\lambda_2-w_2)}_{M_2}(\alpha_a^\dagger)\right)$$
$$= \sum_{k=0}^{w_2}(-1)^k\sqrt{\frac{(\lambda_2+k-w_2+1)!}{(\lambda_2-w_2+1)!}\frac{w_2!}{k!(w_2-k)!}}$$

$$\times\frac{(\alpha_{3a}^{\dagger})^{w_2-k}}{\sqrt{(w_2-k)!}}$$
$$\times\left[P^{(1/2)(\lambda_2-w_2+k)}(\alpha_a^{\dagger})\times Z^{k/2}(\mathbf{z})\right]^{(1/2)(\lambda_2-w_2)}_{M_2}. \tag{34}$$

This is the needed VCS realization of the totally symmetric U(3) tensor.

III. MATRIX ELEMENTS OF INTRINSIC OPERATORS

In order to evaluate matrix elements of the totally symmetric U(3) tensors between states of arbitrary $(\lambda_1\mu_1)$ and $(\lambda_3\mu_3)$,

$$\langle(\lambda_3\mu_3)w_3I_3M_3|T^{(\lambda_2 0)}_{w_2I_2M_2}|(\lambda_1\mu_1)w_1I_1M_1\rangle$$
$$=\langle(\lambda_3\mu_3)w_3I_3\|T^{(\lambda_2 0)}_{w_2I_2}\|(\lambda_1\mu_1)w_1I_1\rangle$$
$$\times\langle I_1M_1I_2M_2|I_3M_3\rangle, \tag{35}$$

via VCS techniques through the use of the basic relation (26) and the $\Gamma(T^{(\lambda_2 0)})$ of Eq. (34) it is necessary to evaluate the matrix elements of the intrinsic operators $(\alpha_{3a}^{\dagger})^{\nu}P^{(1/2)(\lambda_2-\nu)}(\alpha_a^{\dagger})$ in the z-space basis. Note again that these intrinsic operators commute with z_i, $\partial/\partial z_i$. In a z-space matrix element they must be worked through to the left where they are then defined through their matrix elements between a pure intrinsic state on their left and the appropriate permitted states on their right.

It is sufficient to describe the intrinsic states of $(\lambda_1\mu_1)$ by two-rowed tableaux. For the most general coupling $(\lambda_1\mu_1)\times(\lambda_2 0)\to(\lambda_3\mu_3)$ in which n squares are added to row 3 of this tableau [with $0\leqslant n\leqslant\min(\lambda_2,\mu_1)$], the intrinsic states for $(\lambda_3\mu_3)$ with $w_3=0$ will then have only n oscillator excitations of type 3. Since the $(\alpha_{3a}^{\dagger})^{\nu}$ are defined through their left actions on the intrinsic states of $(\lambda_3\mu_3)$, and since the left action of $\alpha_{3a}^{\dagger}$ annihilates an oscillator excitation of type 3, only operators with $\nu\leqslant n$ will have nonzero matrix elements. The only terms of Eq. (34) that can contribute are those with $k=w_2, w_2-1, \ldots, w_2-n$. Note that, with $n=0$ and consequently $\lambda_3+2\mu_3=\lambda_1+2\mu_1+\lambda_2$, the basic relation of Eq. (26) leads to

$$\langle(\lambda_3\mu_3)w_3=0, I_3=(\lambda_3/2)\|T^{(\lambda_2 0)}_{w_2=0,I_2=\lambda_2/2}$$
$$\times\|(\lambda_1\mu_1)w_1=0, I_1=\lambda_1/2\rangle$$
$$=\langle(\lambda_3\mu_3)0(\lambda_3/2)\|P^{\lambda_2/2}(\alpha^{\dagger})\|(\lambda_1\mu_1)0(\lambda_1/2)\rangle$$
$$=\langle(\lambda_3\mu_3)0(\lambda_3/2)\|P^{\lambda_2/2}(\alpha^{\dagger})\|(\lambda_1\mu_1)0(\lambda_1/2)\rangle. \tag{36}$$

In this case the only needed reduced matrix element of intrinsic operators is related immediately to a very simple reduced matrix element in ordinary space. This is the reason for the simplicity of the Le Blanc-Biedenharn result. Moreover, in this case the ordinary space angular momentum reduced matrix element [denoted by standard double lines in Eqs. (35) and (36)] can be reduced to an overall SU(3) reduced matrix element (to be denoted by both double lines and double brackets[9]) through an $SU(3)\supset SU(2)\times U(1)$ Wigner coefficient with value 1. With $\lambda_3+2\mu_3=\lambda_1+2\mu_1+\lambda_2$:

$$\langle(\lambda_3\mu_3)0(\lambda_3/2)\|P^{\lambda_2/2}(\alpha^{\dagger})\|(\lambda_1\mu_1)0(\lambda_1/2)\rangle$$
$$=\langle(\lambda_1\mu_1)Y_1=\tfrac{1}{3}(\lambda_1+2\mu_1)I_1=\lambda_1/2;(\lambda_2 0)\tfrac{1}{3}\lambda_2,\tfrac{1}{2}\lambda_2\|(\lambda_3\mu_3)Y_3=\tfrac{1}{3}(\lambda_3+2\mu_3)I_3=\lambda_3/2\rangle$$
$$\times\langle\langle(\lambda_3\mu_3)\|T^{(\lambda_2 0)}(\alpha^{\dagger})\|(\lambda_1\mu_1)\rangle\rangle=1\times\langle\langle(\lambda_3\mu_3)\|T^{(\lambda_2 0)}(\alpha^{\dagger})\|(\lambda_1\mu_1)\rangle\rangle. \tag{37}$$

Although this $\langle\langle\|\ \|\rangle\rangle$ can be evaluated,[9] it will drop out of all final expressions andis therefore not needed. (Note also that the particle index a on $\alpha^{\dagger}$ has been dropped and will be omitted henceforth for economy of notation, although it is to be quietly understood.)

For the case of arbitrary n, the intrinsic operator matrix elements of operators $(\alpha_3^{\dagger})^{\nu}P^{1/2(\lambda_2-\nu)}(\alpha^{\dagger})$ with $\nu=0,1,\ldots, n$, will be related to the standard Hilbert space matrix element

$$\langle(\lambda_3\mu_3)0(\lambda_3/2)\|(\alpha_3^{\dagger})^{n}P^{(1/2)(\lambda_2-n)}(\alpha^{\dagger})\|(\lambda_1\mu_1)0(\lambda_1/2)\rangle$$

by an inductive process through repeated use of the basic relation (26). The matrix element of the intrinsic operator $(\alpha_3^{\dagger})^{\nu}P^{(1/2)(\lambda_2-\nu)}(\alpha^{\dagger})$ between a purely intrinsic state of $(\lambda_3\mu_3)$ on the left must have $n-\nu$ z excitations in the state of $(\lambda_1\mu_1)$ on the right. This type of matrix element is evaluated by a tranformation back to ordinary Hilbert space, via Eq. (26), where the states with $n-\nu$ excitations are constructed by ordinary Hilbert space creation operator excitations $A_i^{\dagger}$ through Eq. (22). The evaluation of the matrix elements in VCS space of the formal intrinsic operators is thus reduced to the evaluation of ordinary matrix elements in ordinary Hilbert space. The intrinsic operator matrix elements therefore become fully explicit and well defined.

The case $n=1$ will be illustrated in detail. In this case only intrinsic operators with $\nu=0$ and $\nu=1$ lead to nonzero matrix elements. With $w_2=0$, Eq. (34) leads to

$$\Gamma(P^{\lambda_2/2}_{M_2}(\alpha^{\dagger}))=P^{\lambda_2/2}_{M_2}(\alpha^{\dagger}), \tag{38}$$

and the basic relation (26) gives

$$\langle(\lambda_3\mu_3)w_3$$
$$=0(\lambda_3/2)\|P^{\lambda_2/2}(\alpha^{\dagger})\|(\lambda_1\mu_1)w_1=1\quad I_1=(\lambda'/2)\rangle$$
$$=\left((\lambda_3\mu_3)\ \frac{\lambda_3}{2}\|\gamma(P^{\lambda_2/2}(\alpha^{\dagger}))\|(\lambda_1\mu_1)\left[\frac{1}{2}\times\frac{\lambda_1}{2}\right]\frac{\lambda'}{2}\right)$$
$$=1\times\left((\lambda_3\mu_3)\ \frac{\lambda_3}{2}\|P^{\lambda_2/2}(\alpha^{\dagger})\|(\lambda_1\mu_1)\left[\frac{1}{2}\times\frac{\lambda_1}{2}\right]\frac{\lambda'}{2}\right)$$
$$\times K(\lambda_1\mu_1)_{1(\lambda'/2)}, \tag{39}$$

where the z-space states are given by the angular momentum coupled notation, $|(\lambda\mu)[w/2\times\lambda/2]I,M\rangle$, of Eq. (24). Note that the "collective" angular momentum of the intrin-

sic state, $w_3/2 = 0$, is to be omitted in this notation. Note also that the left action of the intrinsic operator $P^{\lambda_2/2}(\partial\!\!\!/^\dagger)$ on the intrinsic state $((\lambda_3\mu_3)(\lambda_3/2)m|$ must create a state with one z-space excitation since $n = 1$ in the space of states with $\lambda_1 + 2\mu_1 = \lambda_3 + 2\mu_3 - \lambda_2 + 3$. By expressing the standard Hilbert space state with $w_1 = 1$ in terms of the $|\phi_{w_1 I_1 M_1}\rangle$ of Eq. (22), the needed intrinsic operator matrix element is given by

$$\left((\lambda_3\mu_3)\frac{\lambda_3}{2}\|P^{\lambda_2/2}(\partial\!\!\!/^\dagger)\|(\lambda_1\mu_1)\left[\frac{1}{2}\times\frac{\lambda_1}{2}\right]\frac{\lambda'}{2}\right)$$
$$= \frac{1}{K^2(\lambda'\mu_1)_{1(\lambda'/2)}}\left\langle(\lambda_3\mu_3)w_3 = 0\frac{\lambda_3}{2}m_3\right|$$
$$\times\left[P^{\lambda_2/2}(\alpha^\dagger)\times\left[\mathbf{A}^{\dagger 1/2}\times|(\lambda_1\mu_1)w_1 = 0\frac{\lambda_1}{2}\right\rangle\right]^{\lambda'/2}\Bigg]^{\lambda_3/2}_{m_3}, \tag{40}$$

where we have used $Z^{1/2}(\mathbf{A}^\dagger) = \mathbf{A}^{\dagger 1/2}$ and the reduced matrix element relation

$$\langle I_3\|T^{I_2}\|I_1\rangle = \langle I_3 M_3|[T^{I_2}|I_1\rangle]^{I_3}_{M_3}. \tag{41}$$

Angular momentum recoupling yields

$$\left((\lambda_3\mu_3)\frac{\lambda_3}{2}\|P^{\lambda_2/2}(\partial\!\!\!/^\dagger)\|(\lambda_1\mu_1)\left[\frac{1}{2}\times\frac{\lambda_1}{2}\right]\frac{\lambda'}{2}\right)$$
$$= \frac{1}{K^2(\lambda_1\mu_1)_{1(\lambda'/2)}}U\left(\frac{\lambda_1}{2}\frac{1}{2}\frac{\lambda_3}{2}\frac{\lambda_2}{2};\frac{\lambda'}{2}\frac{\lambda_2-1}{2}\right)$$
$$\times\left\langle(\lambda_3\mu_3)w_3 = 0\frac{\lambda_3}{2}m_3\right|\Big[[P^{\lambda_2/2}(\alpha^\dagger),A^{\dagger 1/2}]^{(\lambda_2-1)/2}$$
$$\times\left|(\lambda_1\mu_1)w_1 = 0\frac{\lambda_1}{2}\right\rangle\Big]^{\lambda_3}_{m_3}, \tag{42}$$

where the U-coefficient is a Racah coefficient in unitary form, and where the angular momentum-coupled operator $[P^{\lambda_2/2}\times A^{\dagger 1/2}]^{(\lambda_2-1)/2}$ has been converted to an angular momentum-coupled commutator by using the fact that $\mathbf{A}^\dagger$ annihilates the intrinsic state in its left action on this state [cf. Eq. (3)]. The angular momentum-coupled commutator is defined by

$$[P^{\lambda_2/2}(\alpha^\dagger),A^{\dagger 1/2}]^{(\lambda_2-1)/2}_m$$
$$= \sum_{m_1}\left\langle\frac{1}{2}m_1\frac{\lambda_2}{2}m - m_1\middle|\frac{\lambda_2-1}{2}m\right\rangle$$
$$\times[P^{\lambda_2/2}_{m-m_1}(\alpha^\dagger),A^{\dagger 1/2}_{m_1}], \tag{43}$$

with

$$\{A^{\dagger 1/2}_{-1/2}, -A^{\dagger 1/2}_{+1/2}\}, = \{A^\dagger_1,A^\dagger_2\}\equiv\{E_{31},E_{32}\}, \tag{44}$$

this gives

$$[P^{\lambda_2/2}(\alpha^\dagger),A^{\dagger 1/2}]^{(\lambda_2-1)/2}_m = \sqrt{(\lambda_2+1)}\alpha^\dagger_3 P^{(\lambda_2-1)/2}_m(\alpha^\dagger); \tag{45}$$

so that

$$\left((\lambda_3\mu_3)\frac{\lambda_3}{2}\|P^{\lambda_2/2}(\partial\!\!\!/^\dagger)\|(\lambda_1\mu_1)\left[\frac{1}{2}\times\lambda_1\right]\frac{\lambda'}{2}\right)$$
$$= \frac{\sqrt{(\lambda_2+1)}}{K^2(\lambda_1\mu_1)_{1(\lambda'/2)}}U\left(\frac{\lambda_1}{2}\frac{1}{2}\frac{\lambda_3}{2}\frac{\lambda_2}{2};\frac{\lambda'}{2}\frac{\lambda_2-1}{2}\right)$$
$$\times\left\langle(\lambda_3\mu_3)0\frac{\lambda_3}{2}\|\alpha^\dagger_3 P^{(\lambda_2-1)/2}(\alpha^\dagger)\|(\lambda_1\mu_1)0\frac{\lambda_1}{2}\right\rangle, \tag{46}$$

which is the first intrinsic space matrix element needed for $n = 1$.

For the second needed intrinsic matrix element for the case $n = 1$, we use Eq. (34) to give

$$\Gamma(\alpha^\dagger_3 P^{(\lambda_2-1)/2}_m(\alpha^\dagger))$$
$$= \partial\!\!\!/^\dagger_3 P^{(\lambda_2-1)/2}_m(\partial\!\!\!/^\dagger)$$
$$- \sqrt{(\lambda_2+1)}[P^{\lambda_2/2}(\partial\!\!\!/^\dagger)\times Z^{1/2}]^{(\lambda_2-1)/2}_m. \tag{47}$$

The basic relation (26) yields

$$\left\langle(\lambda_3\mu_3)0\frac{\lambda_3}{2}\|\alpha^\dagger_3 P^{(\lambda_2-1)/2}(\alpha^\dagger)\|(\lambda_1\mu_1)0\frac{\lambda_1}{2}\right\rangle - \left((\lambda_3\mu_3)\frac{\lambda_3}{2}\|\partial\!\!\!/^\dagger_3 P^{(\lambda_2-1)/2}(\partial\!\!\!/^\dagger)\|(\lambda_1\mu_1)\frac{\lambda_1}{2}\right)$$
$$= -\sqrt{(\lambda_2+1)}\left((\lambda_3\mu_3)\frac{\lambda_3}{2}\|[P^{\lambda_2/2}(\partial\!\!\!/^\dagger)\,Z^{1/2}]^{(\lambda_2-1)/2}\|(\lambda_1\mu_1)\frac{\lambda_1}{2}\right)$$
$$= -\sqrt{(\lambda_2+1)}\sum_{\lambda'/2}U\left(\frac{\lambda_1}{2}\frac{1}{2}\frac{\lambda_3}{2}\frac{\lambda_2}{2};\frac{\lambda'}{2}\frac{\lambda_2-1}{2}\right)\left((\lambda_3\mu_3)\frac{\lambda_3}{2}\|P^{\lambda_2/2}(\partial\!\!\!/^\dagger)\|(\lambda_1\mu_1)\left[\frac{1}{2}\times\frac{\lambda_1}{2}\right]\frac{\lambda'}{2}\right). \tag{48}$$

The last intrinsic matrix element in Eq. (48) is given by Eq. (46), so that

$$\left((\lambda_3\mu_3)\frac{\lambda_3}{2}\|\partial\!\!\!/^\dagger_3 P^{(\lambda_2-1)/2}(\partial\!\!\!/^\dagger)\|(\lambda_1\mu_1)\frac{\lambda_1}{2}\right)$$
$$= \left\{1 + (\lambda_2+1)\sum_{\lambda'/2}\frac{1}{K^2(\lambda_1\mu_1)_{1(\lambda'/2)}}U^2\left(\frac{\lambda_1}{2}\frac{1}{2}\frac{\lambda_3}{2}\frac{\lambda_2}{2};\frac{\lambda'}{2}\frac{\lambda_2-1}{2}\right)\right\}\left\langle(\lambda_3\mu_3)0\frac{\lambda_3}{2}\|\alpha^\dagger_3 P^{(\lambda_2-1)/2}(\alpha^\dagger)\|(\lambda_1\mu_1)0\frac{\lambda_1}{2}\right\rangle$$
$$= \frac{(\lambda_3+\mu_3+3)(\mu_3+2)}{(\lambda_1+\mu_1+1)\mu_1}\left\langle(\lambda_3\mu_3)0\frac{\lambda_3}{2}\|\alpha^\dagger_3 P^{\lambda_2/2-1/2}(\alpha^\dagger)\|(\lambda_1\mu_1)0\frac{\lambda_1}{2}\right\rangle, \tag{49}$$

where we have used $\lambda_3 + 2\mu_3 = \lambda_1 + 2\mu_1 + \lambda_2 - 3$ to simplify the final result for the sum over possible $\lambda'/2$. Equations (46) and (49) give the two needed intrinsic operator matrix elements for the case $n = 1$.

The case $n = 2$ illustrates some additional features. Three intrinsic operator matrix elements are now needed. The analogs of Eqs. (39)–(46) yield

$$\left((\lambda_3\mu_3)\frac{\lambda_3}{2}\|P^{\lambda_2/2}(\alpha^\dagger)\|(\lambda_1\mu_1)\left[1\times\frac{\lambda_1}{2}\right]\frac{\lambda''}{2}\right)$$
$$=\frac{\sqrt{\lambda_2(\lambda_2+1)}}{K^2(\lambda_1\mu_1)_{2(\lambda''/2)}}U\left(\frac{\lambda_1}{2}1\frac{\lambda_3}{2}\frac{\lambda_2}{2};\frac{\lambda''}{2}\frac{\lambda_2}{2}-1\right)\left\langle(\lambda_3\mu_3)0\frac{\lambda_3}{2}\left|\left|\frac{(\alpha_3^\dagger)^2}{\sqrt{2}}P^{\lambda_2/2-1}(\alpha^\dagger)\right|\right|(\lambda_1\mu_1)0\frac{\lambda_1}{2}\right\rangle, \tag{50}$$

where we have used $Z^1_m(\mathbf{A}^\dagger)=2^{-1/2}[A^{\dagger 1/2}\times A^{\dagger 1/2}]^1_m$, angular momentum recoupling, and a double application of the commutator Eq. (45).

For the next matrix element, Eqs. (26) and (47) yield

$$\left\langle(\lambda_3\mu_3)0\frac{\lambda_3}{2}\|\alpha_3^\dagger P^{(\lambda_2-1)/2}(\alpha^\dagger)\|(\lambda_1\mu_1)w_1=1\frac{\lambda'}{2}\right\rangle$$
$$=\left\{\left((\lambda_3\mu_3)\frac{\lambda_3}{2}\|\alpha_3^\dagger P^{(\lambda_2-1)/2}(\alpha^\dagger)\|(\lambda_1\mu_1)\left[\frac{1}{2}\times\frac{\lambda_1}{2}\right]\frac{\lambda'}{2}\right)\right.$$
$$\left.-\sqrt{(\lambda_2+1)}\left((\lambda_3\mu_3)\frac{\lambda_3}{2}\|[P^{\lambda_2/2}(\alpha^\dagger)\times Z^{1/2}]^{(\lambda_2-1)/2}\|(\lambda_1\mu_1)\left[\frac{1}{2}\times\frac{\lambda_1}{2}\right]\frac{\lambda'}{2}\right)\right\}K(\lambda_1\mu_1)_{1(\lambda'/2)}. \tag{51}$$

The left-hand side is evaluated by the technique used to evaluate the standard Hilbert space matrix element in Eq. (40) to give

$$\left\langle(\lambda_3\mu_3)0\frac{\lambda_3}{2}\|\alpha_3^\dagger P^{(\lambda_2-1)/2}(\alpha^\dagger)\|(\lambda_1\mu_1)w_1=1\frac{\lambda'}{2}\right\rangle$$
$$=\frac{1}{K(\lambda_1\mu_1)_{1(\lambda'/2)}}\sqrt{2\lambda_2}U\left(\frac{\lambda_1}{2}\frac{1}{2}\frac{\lambda_3}{2}\frac{\lambda_2-1}{2};\frac{\lambda'}{2}\frac{\lambda_2}{2}-1\right)\left\langle(\lambda_3\mu_3)0\frac{\lambda_3}{2}\left|\left|\frac{(\alpha_3^\dagger)^2}{\sqrt{2}}P^{\lambda_2/2-1}(\alpha^\dagger)\right|\right|(\lambda_1\mu_1)0\frac{\lambda_1}{2}\right\rangle. \tag{52}$$

Angular momentum recoupling together with Eq. (19) reduces the second term of the right-hand side of Eq. (51) to the intrinsic operator reduced matrix element of $P^{\lambda_2/2}(\alpha^\dagger)$ which is given by Eq. (50), so that

$$\left((\lambda_3\mu_3)\frac{\lambda_3}{2}\|\alpha_3^\dagger P^{(\lambda_2-1)/2}(\alpha^\dagger)\|(\lambda_1\mu_1)\left[\frac{1}{2}\times\frac{\lambda_1}{2}\right]\frac{\lambda'}{2}\right)$$
$$=\left\langle(\lambda_3\mu_3)0\frac{\lambda_3}{2}\left|\left|\frac{(\alpha_3^\dagger)^2}{\sqrt{2}}P^{\lambda_2/2-1}(\alpha^\dagger)\right|\right|(\lambda_1\mu_1)0\frac{\lambda_1}{2}\right\rangle\left\{\frac{\sqrt{2\lambda_2}}{K^2(\lambda_1\mu_1)_{1(\lambda'/2)}}U\left(\frac{\lambda_1}{2}\frac{1}{2}\frac{\lambda_3}{2}\frac{\lambda_2-1}{2};\frac{\lambda'}{2}\frac{\lambda_2}{2}-1\right)\right.$$
$$\left.+\sum_{\lambda''/2}\frac{(\lambda_2+1)\sqrt{2\lambda_2}}{K^2(\lambda_1\mu_1)_{2(\lambda''/2)}}U\left(\frac{\lambda'}{2}\frac{1}{2}\frac{\lambda_3}{2}\frac{\lambda_2}{2};\frac{\lambda''}{2}\frac{\lambda_2-1}{2}\right)U\left(\frac{\lambda_1}{2}\frac{1}{2}\frac{\lambda''}{2}\frac{1}{2};\frac{\lambda'}{2}1\right)U\left(\frac{\lambda_1}{2}1\frac{\lambda_3}{2}\frac{\lambda_2}{2};\frac{\lambda''}{2}\frac{\lambda_2}{2}-1\right)\right\}$$
$$=\left\langle(\lambda_3\mu_3)0\frac{\lambda_3}{2}\left|\left|\frac{(\alpha_3^\dagger)^2}{\sqrt{2}}P^{\lambda_2/2-1}(\alpha^\dagger)\right|\right|(\lambda_1\mu_1)0\frac{\lambda_1}{2}\right\rangle$$
$$\times\frac{\sqrt{2\lambda_2}}{K^2(\lambda_1\mu_1-1)_{1(\lambda'/2)}}U\left(\frac{\lambda_1}{2}\frac{1}{2}\frac{\lambda_3}{2}\frac{\lambda_2-1}{2};\frac{\lambda'}{2}\frac{\lambda_2}{2}-1\right)\frac{(\lambda_3+\mu_3+4)(\mu_3+2)}{(\lambda_1+\mu_1+1)\mu_1}, \tag{53}$$

where we have used $\lambda_1+2\mu_1+\lambda_2-6=\lambda_3+2\mu_3$ to simplify the final result for the sum over possible $\lambda''/2$.

For the final intrinsic operator matrix element for $n=2$, the basic relation (26) gives

$$\left\langle(\lambda_3\mu_3)0\frac{\lambda_3}{2}\left|\left|\frac{(\alpha_3^\dagger)^2}{\sqrt{2}}P^{\lambda_2/2-1}(\alpha^\dagger)\right|\right|(\lambda_1\mu_1)0\frac{\lambda_1}{2}\right\rangle$$
$$=\left((\lambda_3\mu_3)\frac{\lambda_3}{2}\left|\left|\frac{(\alpha_3^\dagger)^2}{\sqrt{2}}P^{\lambda_2/2-1}(\alpha^\dagger)\right|\right|(\lambda_1\mu_1)\frac{\lambda_1}{2}\right)-\sqrt{2\lambda_2}\left((\lambda_3\mu_3)\frac{\lambda_3}{2}\|\alpha_3^\dagger[P^{(\lambda_2-1)/2}(\alpha^\dagger)\times Z^{1/2}]^{\lambda_2/2-1}\|(\lambda_1\mu_1)\frac{\lambda_1}{2}\right)$$
$$+\sqrt{\lambda_2(\lambda_2+1)}\left((\lambda_3\mu_3)\frac{\lambda_3}{2}\|[P^{\lambda_2/2}(\alpha^\dagger)\times Z^1]^{\lambda_2/2-1}\|(\lambda_1\mu_1)\frac{\lambda_1}{2}\right). \tag{54}$$

Angular momentum recoupling reduces the last two terms to the form that follows from Eqs. (50) and (53), so that

$$\left((\lambda_3\mu_3)\frac{\lambda_3}{2}\left|\left|\frac{(\alpha_3^\dagger)^2}{\sqrt{2}}P(\alpha^\dagger)^{\lambda_2/2-1}\right|\right|(\lambda_1\mu_1)\frac{\lambda_1}{2}\right)$$
$$=\left\langle(\lambda_3\mu_3)0\frac{\lambda_3}{2}\left|\left|\frac{(\alpha_3^\dagger)^2}{\sqrt{2}}P(\alpha^\dagger)^{\lambda_2/2-1}\right|\right|(\lambda_1\mu_1)0\frac{\lambda_1}{2}\right\rangle\left\{1+2\lambda_2\sum_{\lambda'/2}\frac{1}{K^2(\lambda_1\mu_1)_{1(\lambda'/2)}}U^2\left(\frac{\lambda_1}{2}\frac{1}{2}\frac{\lambda_3}{2}\frac{\lambda_2-1}{2};\frac{\lambda'}{2}\frac{\lambda_2}{2}-1\right)\right.$$
$$\left.+2\lambda_2(\lambda_2+1)\sum_{\lambda''/2}\left[\sum_{\lambda'/2}U\left(\frac{\lambda_1}{2}\frac{1}{2}\frac{\lambda_3}{2}\frac{\lambda_2-1}{2};\frac{\lambda'}{2}\frac{\lambda_2}{2}-1\right)U\left(\frac{\lambda'}{2}\frac{1}{2}\frac{\lambda_3}{2}\frac{\lambda_2}{2};\frac{\lambda''}{2}\frac{\lambda_2-1}{2}\right)\right.\right.$$

$$U\left(\frac{\lambda_1}{2}\,\frac{1}{2}\,\frac{\lambda''}{2}\,\frac{1}{2};\frac{\lambda'}{2}\,1\right)\Big]\frac{1}{K^2(\lambda_1\mu_1)_{2(\lambda''/2)}}U\left(\frac{\lambda_1}{2}\,1\,\frac{\lambda_3}{2}\,\frac{\lambda_2}{2};\frac{\lambda''}{2}\,\frac{\lambda_2}{2}-1\right)$$

$$\times -\lambda_2(\lambda_2+1)\sum_{\lambda''/2}\frac{1}{K^2(\lambda_1\mu_1)_{2(\lambda''/2)}}U^2\left(\frac{\lambda_1}{2}\,1\,\frac{\lambda_3}{2}\,\frac{\lambda_2}{2};\frac{\lambda''}{2}\,\frac{\lambda_2}{2}-1\right)\Big]. \tag{55}$$

It is now advantageous to carry out the sum over $\lambda'/2$ first in the third term, where

$$\sum_{\lambda''/2}U\left(\frac{\lambda_1}{2}\,\frac{1}{2}\,\frac{\lambda_3}{2}\,\frac{\lambda_2-1}{2};\frac{\lambda'}{2}\,\frac{\lambda_2}{2}-1\right)U\left(\frac{\lambda'}{2}\,\frac{1}{2}\,\frac{\lambda_3}{2}\,\frac{\lambda_2}{2};\frac{\lambda''}{2}\,\frac{\lambda_2-1}{2}\right)\times U\left(\frac{\lambda_1}{2}\,\frac{1}{2}\,\frac{\lambda''}{2}\,\frac{1}{2};\frac{\lambda'}{2}\,1\right)=U\left(\frac{\lambda_1}{2}\,1\,\frac{\lambda_3}{2}\,\frac{\lambda_2}{2};\frac{\lambda''}{2}\,\frac{\lambda_2}{2}-1\right), \tag{56}$$

to obtain

$$\left((\lambda_3\mu_3)\,\frac{\lambda_3}{2}\,\middle|\middle|\,\frac{(\alpha_3^\dagger)^2}{\sqrt{2}}P^{\lambda_3/2-1}(\alpha^\dagger)\,\middle|\middle|\,(\lambda_1\mu_1)\,\frac{\lambda_1}{2}\right)$$

$$=\left((\lambda_3\mu_3)0\,\frac{\lambda_3}{2}\,\middle|\middle|\,\frac{(\alpha_3^\dagger)^2}{\sqrt{2}}P^{\lambda_3/2-1}(\alpha^\dagger)\,\middle|\middle|\,(\lambda_1\mu_1)0\,\frac{\lambda_1}{2}\right)$$

$$\Big\{1+2\lambda_2\sum_{\lambda'/2}\frac{1}{K^2(\lambda_1\mu_1)_{1(\lambda'/2)}}U^2\left(\frac{\lambda_1}{2}\,\frac{1}{2}\,\frac{\lambda_3}{2}\,\frac{\lambda_2-1}{2};\frac{\lambda'}{2}\,\frac{\lambda_2}{2}-1\right)$$

$$\lambda_2(\lambda_2+1)\sum_{\lambda''/2}\frac{1}{K^2(\lambda_1\mu_1)_{2(\lambda''/2)}}U^2\left(\frac{\lambda_1}{2}\,1\,\frac{\lambda_3}{2}\,\frac{\lambda_2}{2};\frac{\lambda''}{2}\,\frac{\lambda_2}{2}-1\right)\Big\}$$

$$=\left((\lambda_3\mu_3)0\,\frac{\lambda_3}{2}\,\middle|\middle|\,\frac{(\alpha_3^\dagger)^2}{\sqrt{2}}P^{\lambda_3/2-1}(\alpha^\dagger)\,\middle|\middle|\,(\lambda_1\mu_1)0\,\frac{\lambda_1}{2}\right)\frac{(\lambda_3+\mu_3+4)(\lambda_3+\mu_3+3)(\mu_3+3)(\mu_3+2)}{(\lambda_1+\mu_1+1)(\lambda_1+\mu_1)\mu_1(\mu_1-1)}. \tag{57}$$

For the case of general n (with $\lambda_3+2\mu_3=\lambda_1+2\mu_1+\lambda_2-3n$), the corresponding result is

$$\left((\lambda_3\mu_3)\,\frac{\lambda_3}{2}\,\middle|\middle|\,\frac{(\alpha_3^\dagger)^n}{\sqrt{n!}}P^{(\lambda_2-n)/2}(\alpha^\dagger)\,\middle|\middle|\,(\lambda_1\mu_1)\,\frac{\lambda_1}{2}\right)$$

$$=\left((\lambda_3\mu_3)0\,\frac{\lambda_3}{2}\,\middle|\middle|\,\frac{(\alpha_3^\dagger)^n}{\sqrt{n!}}P^{(\lambda_2-n)/2}(\alpha^\dagger)\,\middle|\middle|\,(\lambda_1\mu_1)0\,\frac{\lambda_1}{2}\right)\sum_{\nu=0}^{n}\sum_{\lambda''/2}\frac{(\lambda_2+1-\nu)!}{(\lambda_2+1-n)!}\frac{n!}{(n-\nu)!\nu!}\frac{1}{K^2(\lambda_1\mu_1)_{(n-\nu)(\lambda''/2)}}$$

$$\times U^2\left(\frac{\lambda_1}{2}\,\frac{n-\nu}{2}\,\frac{\lambda_3}{2}\,\frac{\lambda_2-\nu}{2};\frac{\lambda''}{2}\,\frac{\lambda_2-n}{2}\right)$$

$$=\left((\lambda_3\mu_3)0\,\frac{\lambda_3}{2}\,\middle|\middle|\,\frac{(\alpha_3^\dagger)^n}{\sqrt{n!}}P^{(\lambda_2-n)/2}(\alpha^\dagger)\,\middle|\middle|\,(\lambda_1\mu_1)0\,\frac{\lambda_1}{2}\right)$$

$$\times\frac{(\lambda_3+\mu_3+n+2)!(\mu_3+n+1)!(\lambda_1+\mu_1+1-n)!(\mu_1-n)!}{(\lambda_3+\mu_3+2)!(\mu_3+1)!(\lambda_1+\mu_1+1)!\mu_1!}. \tag{58}$$

The techniques illustrated in detail for $n=1$ and $n=2$ lead to the general result for the needed intrinsic operator matrix elements

$$\left((\lambda_3\mu_3)\,\frac{\lambda_3}{2}\,\middle|\middle|\,\frac{(\alpha_3^\dagger)^\nu}{\sqrt{\nu!}}P^{(\lambda_2-\nu)/2}(\alpha^\dagger)\,\middle|\middle|\,(\lambda_1\mu_1)\left[\frac{n-\nu}{2}\times\frac{\lambda_1}{2}\right]\frac{\lambda''}{2}\right)$$

$$=\left((\lambda_3\mu_3)0\,\frac{\lambda_3}{2}\,\middle|\middle|\,\frac{(\alpha_3^\dagger)^n}{\sqrt{n!}}P^{(\lambda_2-n)/2}(\alpha^\dagger)\,\middle|\middle|\,(\lambda_1\mu_1)0\,\frac{\lambda_1}{2}\right)\sqrt{\frac{(\lambda_2+1-\nu)!}{(\lambda_2+1-n)!}\frac{n!}{(n-\nu)!\nu!}}\frac{1}{K^2(\lambda_1,\mu_1-\nu)_{(n-\nu)\lambda''/2}}$$

$$\times U\left(\frac{\lambda_1}{2}\,\frac{n-\nu}{2}\,\frac{\lambda_3}{2}\,\frac{\lambda_2-\nu}{2};\frac{\lambda''}{2}\,\frac{\lambda_2-n}{2}\right)\frac{(\lambda_3+\mu_3+n+2)!(\mu_3+n+1)!(\lambda_1+\mu_1+1-\nu)!(\mu_1-\nu)!}{(\lambda_3+\mu_3+n-\nu+2)!(\mu_3+n-\nu+1)!(\lambda_1+\mu_1+1)!\mu_1!}. \tag{59}$$

Note that Eq. (59) reduces properly to Eq. (58) for $n=\nu$.

Finally, there remains the evaluation of the standard Hilbert space matrix element, with $\nu=n$. This is given by

$$\left((\lambda_3\mu_3)0\,\frac{\lambda_3}{2}\,\middle|\middle|\,\frac{(\alpha_3^\dagger)^n}{\sqrt{n!}}P^{(\lambda_2-n)/2}(\alpha^\dagger)\,\middle|\middle|\,(\lambda_1\mu_1)0\,\frac{\lambda_1}{2}\right)$$

$$=\left((\lambda_1\mu_1)\,Y_1=\frac{1}{3}(\lambda_1+2\mu_1)I_1=\frac{\lambda_1}{2};(\lambda_2 0)\,\frac{1}{3}\lambda_2-n,\frac{1}{2}\lambda_2-\frac{n}{2}\,\middle\|\,(\lambda_3\mu_3)\,Y_3=\frac{1}{3}(\lambda_3+2\mu_3)I_3=\frac{\lambda_3}{2}\right)$$

$$\times\langle\langle(\lambda_3\mu_3)\|T^{(\lambda_2 0)}(\alpha^\dagger)\|(\lambda_1\mu_1)\rangle\rangle. \tag{60}$$

Since the double-bar, double-bracket factor will not be needed, it suffices to evaluate the $SU(3)\supset SU(2)\times U(1)$ reduced

Wigner coefficient with both $w_3 = 0$, $w_1 = 0$. In Sec. IV it will be shown that the square of this Wigner coefficient is given by the inverse of the double sum of Eq. (58). With a generalized Condon–Shortley phase convention, we therefore get, with $\lambda_3 + 2\mu_3 = \lambda_1 + 2\mu_1 + \lambda_2 - 3n$:

$$\left\langle (\lambda_1\mu_1) Y_1 = \frac{1}{3}(\lambda_1 + 2\mu_1) I_1 = \frac{\lambda_1}{2};\quad (\lambda_2 0)\,\frac{1}{3}\lambda_2 - n,\quad \frac{\lambda_2 - n}{2}\middle|\middle| (\lambda_3\mu_3) Y_3 = \frac{1}{3}(\lambda_3 + 2\mu_3) I_3 = \frac{\lambda_3}{2}\right\rangle$$
$$= \sqrt{\frac{(\lambda_3 + \mu_3 + 2)!(\mu_3 + 1)!(\lambda_1 + \mu_1 + 1)!\mu_1!}{(\lambda_3 + \mu_3 + n + 2)!(\mu_3 + n + 1)!(\lambda_1 + \mu_1 + 1 - n)!(\mu_1 - n)!}}. \tag{61}$$

IV. THE SU(3)⊃SU(2)×U(1) WIGNER COEFFICIENT. FORM I

The general $SU(3) \supset SU(2) \times U(1)$ reduced Wigner coefficient for the product $(\lambda_1\mu_1) \times (\lambda_2 0) \to (\lambda_3\mu_3)$ follows from the general matrix element

$$\left\langle (\lambda_3\mu_3) w_3 I_3 \middle|\middle| \frac{\alpha_3^{\dagger w_2}}{\sqrt{w_2!}} P^{(\lambda_2 - w_2)/2}(\alpha^\dagger) \middle|\middle| (\lambda_1\mu_1) w_1 I_1 \right\rangle$$
$$= \left\langle (\lambda_1\mu_1) Y_1 = \frac{1}{3}(\lambda_1 + 2\mu_1) - w_1 I_1; (\lambda_2 0)\,\frac{1}{3}\lambda_2 - w_2, \frac{\lambda_2 - \omega_2}{2}\middle|\middle| (\lambda_3\mu_3) Y_3 \right.$$
$$\left. = \frac{1}{3}(\lambda_3 + 2\mu_3) - w_3 I_3 \right\rangle \ll (\lambda_3\mu_3) \| T^{(\lambda_2 0)} \| (\lambda_1\mu_1) \gg, \tag{62}$$

with $w_3 = w_1 + w_2 - n$ for the coupling with n squares added to row 3 so that $\lambda_3 + 2\mu_3 = \lambda_1 + 2\mu_1 + \lambda_2 - 3n$. The basic relation (26) together with the VCS realization of the totally symmetric tensor, Eq. (34), gives

$$\left\langle (\lambda_3\mu_3) w_3 I_3 \middle|\middle| \frac{(\alpha_3^\dagger)^{w_2}}{\sqrt{w_2!}} P^{(\lambda_2 - w_2)/2}(\alpha^\dagger) \middle|\middle| (\lambda_1\mu_1) w_1 I_1 \right\rangle$$
$$= \frac{K(\lambda_1\mu_1)_{w_1 I_1}}{K(\lambda_3\mu_3)_{w_3 I_3}} \sum_{\nu=0}^{\min(n, w_2)} (-1)^{w_2 - \nu} \sqrt{\frac{(\lambda_2 - \nu + 1)!}{(\lambda_2 - w_2 + 1)!}\frac{w_2!}{(w_2 - \nu)!\nu!}}$$
$$\times \left((\lambda_3\mu_3) \left[\frac{w_3}{2} \times \frac{\lambda_3}{2}\right] I_3 \middle|\middle| \frac{(\alpha_3^\dagger)^\nu}{\sqrt{\nu!}} [P^{(\lambda_2 - \nu)/2}(\alpha^\dagger) Z^{(w_2 - \nu)/2}(\mathbf{z})]^{I_2 = (\lambda_2 - w_2)/2} \middle|\middle| (\lambda_1\mu_1) \left[\frac{w_1}{2} \times \frac{\lambda_1}{2}\right] I_1 \right). \tag{63}$$

The reduced matrix element in Eq. (63), rhs, can be evaluated directly by the expansion

$$\left((\lambda_3\mu_3) \left[\frac{w_3}{2} \times \frac{\lambda_3}{2}\right] I_3 \middle|\middle| \frac{(\alpha_3^\dagger)^\nu}{\sqrt{\nu!}} [P^{(\lambda_2 - \nu)/2}(\alpha^\dagger) \times Z^{(w_2 - \nu)/2}(\mathbf{z})]^{I_2 = (\lambda_2 - \omega_2)/2} \middle|\middle| (\lambda_1\mu_1) \left[\frac{w_1}{2} \times \frac{\lambda_1}{2}\right] I_1 \right)$$
$$= \frac{1}{(2I_3 + 1)} \sum_{\text{all } M_i m_i m_{w_i}} \langle I_1 M_1 I_2 M_2 | I_3 M_3 \rangle \left\langle \frac{\lambda_1}{2} m_1 \frac{w_1}{2} m_{w_1} | I_1 M_1 \right\rangle \left\langle \frac{\lambda_2 - \nu}{2} m_2 \frac{w_2 - n}{2} m_{w_2} \middle| I_2 M_2 \right\rangle$$
$$\times \left\langle \frac{\lambda_3}{2} m_3 \frac{w_3}{2} m_{w_3} | I_3 M_3 \right\rangle \left(\frac{\lambda_3}{2} m_3 \middle| \frac{(\alpha_3^\dagger)^\nu}{\sqrt{\nu!}} P_{m_2}^{(\lambda_2 - \nu)/2}(\alpha^\dagger) (Z_{w_{m_1}}^{w_1/2}(\mathbf{z}))^\dagger Z_{m_{w_2}}^{(\omega_2 - \nu)/2}(\mathbf{z}) Z_{m_{w_1}}^{w_1/2}(\mathbf{z}) \middle| \frac{\lambda_1}{2} m_1 \right), \tag{64}$$

and by using

$$\left(\frac{\lambda_3}{2} m_3 \middle| \frac{(\alpha_3^\dagger)^\nu}{\sqrt{\nu!}} P_{m_2}^{(\lambda_2 - \nu)/2}(\alpha^\dagger) \right.$$
$$= \sum_{\lambda' / 2, m'} \left(\frac{\lambda_3}{2} m_3 \middle| \frac{(\alpha_3^\dagger)^\nu}{\sqrt{\nu!}} P_{m_2}^{(\lambda_2 - \nu)/2}(\alpha^\dagger) \middle| (\lambda_1\mu_1) \left[\frac{n - \nu}{2} \times \frac{\lambda_1}{2}\right] \frac{\lambda'}{2} m' \right) \left((\lambda_1\mu_1) \left[\frac{n - \nu}{2} \times \middle| \frac{\lambda_1}{2} \middle| \frac{\lambda'}{2} m' \right.\right.$$
$$= \sum_{\lambda' / 2} \sum_{m' m_1' m_{n - \nu}} \left((\lambda_3\mu_3) \frac{\lambda_3}{2} \middle|\middle| \frac{(\alpha_3^\dagger)^\nu}{\sqrt{\nu!}} P^{(\lambda_2 - \nu)/2}(\alpha^\dagger) \middle|\middle| (\lambda_1\mu_1) \left[\frac{n - \nu}{2} \times \frac{\lambda_1}{2}\right] \frac{\lambda'}{2} \right)$$
$$\times \left\langle \frac{\lambda'}{2} m' \frac{\lambda_2 - \nu}{2} m_2 \middle| \frac{\lambda_3}{2} m_3 \right\rangle \left\langle \frac{\lambda_1}{2} m_1' \frac{n - \nu}{2} m_{n - \nu} \middle| \frac{\lambda'}{2} m' \right\rangle \left(\frac{\lambda_1}{2} m_1' \middle| (Z_{m_{n - \nu}}^{(n - \nu)/2}(\mathbf{z}))^\dagger \right). \tag{65}$$

Finally, using Eq. (19) and the orthonormality of the $Z_m^{w/2}(\mathbf{z})$ in z space and the $|\lambda_1/2 m_1)$ in intrinsic space

$$\left(\frac{\lambda_1}{2} m_1' \middle| (Z_{m_{n - \nu}}^{(n - \nu)/2}(\mathbf{z}))^\dagger (Z_{m_{w_1}}^{w_1/2}(\mathbf{z}))^\dagger Z_{m_{w_2}}^{(w_2 - \nu)/2}(\mathbf{z}) Z_{m_{w_1}}^{w_1/2}(\mathbf{z}) \middle| \frac{\lambda_1}{2} m_1 \right)$$
$$= \sqrt{\frac{(w_3 + n - \nu)!}{w_3!(n - \nu)!}\frac{(w_1 + w_2 - \nu)!}{w_1!(w_2 - \nu)!}} \left\langle \frac{w_1}{2} m_{w_1} \frac{w_2 - \nu}{2} m_{w_2} \middle| \frac{w_1 + w_2 - \nu}{2} (m_{w_1} + m_{w_2}) \right\rangle$$

$$\times\left\langle \frac{w_3}{2}\, m_{w_3}\ \frac{n-\nu}{2}\, m_{n-\nu}\ \middle|\ \frac{w_3+n-\nu}{2}\,(m_{w_1}+m_{n-\nu})\right\rangle \delta_{m_1' m_1}\,\delta_{(m_{w_1}+m_{n-\nu})(m_{w_1}+m_{w_2})}, \tag{66}$$

with $w_3 = w_1 + w_2 - n$, Eqs. (64)–(66) give

$$\left((\lambda_3\mu_3)\left[\frac{w_3}{2}\times\frac{\lambda_3}{2}\right]I_3\ \middle|\middle|\ \frac{(\alpha_3^\dagger)^\nu}{\sqrt{\nu!}}\,[P^{(\lambda_2-\nu)/2}(\alpha^\dagger)\times Z^{(w_2-\nu)/2}(z)]I_2 = \frac{\lambda_2-w_2}{2}\ \middle|\middle|\ (\lambda_1\mu_1)\left[\frac{w_1}{2}\times\frac{\lambda_1}{2}\right]I_1\right)$$

$$= \sum_{\lambda'/2}\left((\lambda_3\mu_3)\,\frac{\lambda_3}{2}\ \middle|\middle|\ \frac{(\alpha_3^\dagger)^\nu}{\sqrt{\nu!}}\,P^{(\lambda_2-\nu)/2}(\alpha^\dagger)\ \middle|\middle|\ (\lambda_1\mu_1)\left[\frac{n-\nu}{2}\times\frac{\lambda_1}{2}\right]\frac{\lambda'}{2}\right)\frac{(w_1+w_2-\nu)!}{\sqrt{(w_1+w_2-n)!(n-\nu)!(w_2-\nu)!w_1!}}\left\{\Sigma\right\}, \tag{67}$$

where $\{\Sigma\}$ is shorthand for the sum over the product of eight angular momentum Wigner coefficients that can be expressed in terms of a 12-j recoupling coefficient in unitary form:

$$\left\{\Sigma\right\} = \frac{1}{(2I_3+1)}\sum_{\text{all } m\text{'s}}\langle I_1M_1I_2M_2|I_3M_3\rangle$$

$$\times\left\langle\frac{\lambda_1}{2}\,m_1\,\frac{w_1}{2}\,m_{w_1}\middle|I_1M_1\right\rangle\left\langle\frac{\lambda_2-\nu}{2}\,m_2\,\frac{w_2-\nu}{2}\,m_{w_2}\middle|I_2M_2\right\rangle\left\langle\frac{\lambda_3}{2}\,m_3\,\frac{w_3}{2}\,m_{w_3}\middle|I_3M_3\right\rangle\left\langle\frac{\lambda'}{2}\,m'\,\frac{\lambda_2-\nu}{2}\,m_2\middle|\frac{\lambda_3}{2}\,m_3\right\rangle$$

$$\times\left\langle\frac{\lambda_1}{2}\,m_1\,\frac{n-\nu}{2}\,m_{n-\nu}\middle|\frac{\lambda'}{2}\,m'\right\rangle\left\langle\frac{w_1}{2}\,m_{w_1}\,\frac{w_2-\nu}{2}\,m_{w_2}\middle|\frac{w_1+w_2-\nu}{2}\,(m_{w_1}+m_{w_2})\right\rangle$$

$$\times\left\langle\frac{w_3}{2}\,m_{w_3}\,\frac{n-\nu}{2}\,m_{n-\nu}\middle|\frac{w_3+n-\nu}{2}\,(m_{w_1}+m_{w_2})\right\rangle$$

$$= \sqrt{\frac{(2I_1+1)(2I_2+1)(w_1+w_2-\nu+1)(\lambda_3+1)}{(2I_3+1)(\lambda_1+1)(\lambda_2-\nu+1)(w_3+1)}}\,(-1)^{\lambda_1/2-\lambda'/2-\nu/2-n/2+w_2}$$

$$\times\begin{bmatrix} I_2 & I_1 & I_3 & \frac{w_3}{2} \\ \frac{w_2-\nu}{2} & \frac{w_1}{2} & \frac{w_1+w_2-\nu}{2} & \frac{\lambda'}{2} \\ \frac{\lambda_2-\nu}{2} & \frac{\lambda_1}{2} & \frac{\lambda_3}{2} & \frac{n-\nu}{2}\end{bmatrix}$$

$$= \sum_{I''}\sqrt{\frac{(2I''+1)(\lambda_3+1)}{(2I_3+1)(\lambda'+1)}}\,U\left(I_1\,\frac{w_2-\nu}{2}\,I_3\,\frac{\lambda_2-\nu}{2};I''\,\frac{\lambda_2-w_2}{2}\right)U\left(\frac{\lambda_2-\nu}{2}\,\frac{\lambda_3}{2}\,I''\,\frac{w_3}{2};\frac{\lambda'}{2}\,I_3\right)$$

$$\times U\left(\frac{\lambda_1}{2}\,\frac{w_1}{2}\,I''\,\frac{w_2-\nu}{2};I_1\,\frac{w_1+w_2-\nu}{2}\right)U\left(\frac{\lambda_1}{2}\,\frac{n-\nu}{2}\,I''\,\frac{w_3}{2};\frac{\lambda'}{2}\,\frac{w_1+w_2-\nu}{2}\right), \tag{68}$$

where the 12-j coefficient in unitary (square bracket) form is the standard transformation coefficient[19,20]

$$\begin{bmatrix} j_a & j_b & J_{ab} & J' \\ j_c & j_d & J_{cd} & J'' \\ J_{ac} & J_{bd} & J_r & J\end{bmatrix}$$

$$= \langle[[J_r\times[j_a\times j_b]J_{ab}]J'\times[j_c\times j_d]J_{cd}]JM\,|\,[[J_r\times[j_a\times j_c]J_{ac}]J''\times[j_b\times j_d]J_{bd}]JM\rangle, \tag{69}$$

where the square brackets denote angular momentum coupling, for now in the conventional left to right coupling order.

Since the intrinsic operator reduced matrix elements needed for Eq. (67) are known from Eq. (59); Eqs. (63) and (67) can be combined to give the desired $SU(3)\supset SU(2)\times U(1)$ Wigner coefficient in terms of the matrix elements of $(\alpha_3^\dagger)^n P^{(\lambda_2-n)/2}(\alpha^\dagger)$ between extremal states with $w_1 = w_3 = 0$, where the double-line, double-bracket reduced matrix element is eliminated via Eqs. (60) and (62):

$$\frac{\langle(\lambda_1\mu_1)Y_1(w_1)I_1;(\lambda_20)Y_2=\tfrac{1}{3}\lambda_2-w_2,I_2=(\lambda_2-w_2)/2\|(\lambda_3\mu_3)Y_3(w_3)I_3\rangle}{\langle(\lambda_1\mu_1)\tfrac{1}{3}(\lambda_1+2\mu_1)(\lambda_1/2);(\lambda_20)\tfrac{1}{3}\lambda_2-n,(\lambda_2-n)/2\|(\lambda_3\mu_3)\tfrac{1}{3}(\lambda_3+2\mu_3)(\lambda_3/2)\rangle}$$

$$= \sum_{\nu=0}^{\min(n,w_2)}(-1)^\nu\frac{K(\lambda_1\mu_1)_{w_1I_1}}{K(\lambda_3\mu_3)_{w_3I_3}}\sqrt{\frac{(\lambda_2-\nu+1)!w_2!(\lambda_2+1-\nu)!n!}{(\lambda_2-w_2+1)!\nu!(w_2-\nu)!(\lambda_2+1-n)!(n-\nu)!\nu!}}$$

$$\times\frac{(w_1+\omega_2-\nu)!}{\sqrt{(n-\nu)!w_1!(w_2-\nu)!(w_1+w_2-n)!}}$$

$$\times\sum_{\lambda'/2}\frac{1}{K^2(\lambda_1,\mu_1-\nu)_{(n-\nu)(\lambda'/2)}}U\left(\frac{\lambda_1}{2}\,\frac{n-\nu}{2}\,\frac{\lambda_3}{2}\,\frac{\lambda_2-\nu}{2};\frac{\lambda'}{2}\,\frac{\lambda_2-n}{2}\right)$$

$$\times\begin{bmatrix} I_2 & I_1 & I_3 & \frac{w_3}{2} \\ \frac{w_2-\nu}{2} & \frac{w_1}{2} & \frac{w_1+w_2-\nu}{2} & \frac{\lambda'}{2} \\ \frac{\lambda_2-\nu}{2} & \frac{\lambda_1}{2} & \frac{\lambda_3}{2} & \frac{n-\nu}{2} \end{bmatrix}$$

$$\times\sqrt{\frac{(2I_1+1)(2I_2+1)(\lambda_3+1)(w_1+w_2-\nu+1)}{(2I_3+1)(\lambda_1+1)(\lambda_2-\nu+1)(w_3+1)}}\,(-1)^{\lambda_1/2-\lambda'/2-n/2-\nu/2}$$

$$\times\frac{(\mu_3+n+1)!(\lambda_3+\mu_3+n+2)!(\mu_1-\nu)!(\lambda_1+\mu_1+1-\nu)!}{(\mu_3+n-\nu+1)!(\lambda_3+\mu_3+n-\nu+2)!\mu_1!(\lambda_1+\mu_1+1)!}. \tag{70}$$

Setting $w_3=0$, we obtain a relation between the Wigner coefficient with $w_3=0$ but arbitrary $w_1 I_1 (w_2=n-w_1)$ and the starting coefficient [Eq. (61)], with both $w_1=0, w_3=0, (w_2=n)$. Using relations such as those illustrated by Eqs. (56) and the sum over $\lambda''/2$ in Eq. (53), the orthonormality sum over w_1 and I_1 can be put into the form

$$\sum_{w_1,w_2 I_1}\left\langle (\lambda_1\mu_1)Y_1(w_1)I_1;(\lambda_2\mu_2)\frac{\lambda_2}{3}-w_2\,\frac{\lambda_2-w_2}{2}\,||\,(\lambda_3\mu_3)Y_3(0)\,\frac{\lambda_3}{2}\right\rangle^2=1$$

$$=\left\langle (\lambda_1\mu_1)\frac{1}{3}(\lambda_1+2\mu_1)\frac{\lambda_1}{2};(\lambda_2\mu_2)\frac{\lambda_2}{3}-n\,\frac{\lambda_2-n}{2}\,\middle|\,\middle|\,(\lambda_3\mu_3)\frac{1}{3}(\lambda_3+2\mu_3)\frac{\lambda_3}{2}\right\rangle^2$$

$$\times\sum_{\nu=0}^{n}\left\{\frac{(\lambda_2+1-\nu)!n!}{(\lambda_2+1-n)!(n-\nu)!\nu!}\sum_{I_1}\frac{1}{K^2(\lambda_1\mu_1)_{(n-\nu)I_1}}U^2\left(\frac{\lambda_1}{2}\,\frac{n-\nu}{2}\,\frac{\lambda_3}{2}\,\frac{\lambda_2}{2};I_1\,\frac{\lambda_2-n}{2}\right)\right\}. \tag{71}$$

The sum is the same as that which has been evaluated in Eq. (58) and leads at once to the needed starting Wigner coefficient with both $w_1=0$, $w_3=0$, as quoted in Eq. (61). It is interesting to note that the square of this starting Wigner coefficient is given by the ratio of the reduced matrix element of $(\alpha_3^\dagger)^n P^{(\lambda_2-n)/2}(\alpha^\dagger)$ between ordinary Hilbert space states with both $w_1=0$, $w_3=0$ to that of the intrinsic operator $(\mathcal{A}_3^\dagger)^n \mathbf{P}^{(\lambda_1-n)/2}(\mathcal{A}^\dagger)$ between pure intrinsic states.

With the evaluation of this starting Wigner coefficient, the general Wigner coefficient is given by

with $\lambda_3+2\mu_3=\lambda_1+2\mu_1+\lambda_2-3n$, $\quad w_3=w_1+w_2-n$,

Form I:

$$\langle (\lambda_1\mu_1)Y_1(w_1)I_1;(\lambda_2 0)Y_2(w_2)I_2=(\lambda_2-w_2)/2\|(\lambda_3\mu_3)Y_3(w_3)I_3\rangle$$

$$=\frac{K(\lambda_1\mu_1)_{w_1 I_1}}{K(\lambda_3\mu_3)_{w_3 I_3}}\sqrt{\frac{w_3!(\lambda_2+1-n)!(\lambda_3+\mu_3+2)!(\mu_3+1)!(\mu_3+n+1)!(\lambda_3+\mu_3+n+2)!}{w_1!w_2!n!(\lambda_2+1-w_2)!(\lambda_1+\mu_1+1)!\mu_1!(\mu_1-n)!(\lambda_1+\mu_1+1-n)!}}$$

$$\times\sum_{\nu=0}^{\min(n,w_2)}(-1)^\nu\frac{w_2!(\lambda_2+1-\nu)!(w_1+w_2-\nu)!n!(\mu_1-\nu)!(\lambda_1+\mu_1+1-\nu)!}{(w_2-\nu)!(\lambda_2+1-n)!w_3!(n-\nu)!\nu!(\lambda_3+\mu_3+2+n-\nu)!(\mu_3+n+1-\nu)!}$$

$$\times\sum_{\lambda'/2}\frac{1}{K^2(\lambda_1,\mu_1-\nu)_{(n-\nu)\lambda'/2}}U\left(\frac{\lambda_1}{2}\,\frac{n-\nu}{2}\,\frac{\lambda_3}{2}\,\frac{\lambda_2-\nu}{2};\frac{\lambda'}{2}\,\frac{\lambda_2-n}{2}\right)$$

$$\times\begin{bmatrix} I_2 & I_1 & I_3 & \frac{w_3}{2} \\ \frac{w_2-\nu}{2} & \frac{w_1}{2} & \frac{w_1+w_2-\nu}{2} & \frac{\lambda'}{2} \\ \frac{\lambda_2-\nu}{2} & \frac{\lambda_1}{2} & \frac{\lambda_3}{2} & \frac{n-\nu}{2} \end{bmatrix}(-1)^{\lambda_1/2-\lambda'/2-\nu/2-n/2}$$

$$\times\sqrt{\frac{(2I_1+1)(2I_2+1)(w_1+w_2-\nu+1)(\lambda_3+1)}{(2I_3+1)(\lambda_1+1)(\lambda_2-\nu+1)(w_3+1)}}. \tag{72}$$

This form is particularly economical, if n or w_2 are small so that the number of terms in the ν sum is small. In particular, for $n=0$, the sums collapse to a single term, the 12-j coefficient collapses to a 9-j coefficient, and Eq. (72) reduces to the first special Le Blanc–Biedenharn result.

With $n=0$, $\lambda_3+2\mu_3=\lambda_1+2\mu_1+\lambda_2$, $w_3=w_1+w_2$,

$$\langle (\lambda_1\mu_1)Y_1(w_1)I_1;(\lambda_2 0)Y_2(w_2)I_2$$

$$=\frac{\lambda_2-w_2}{2}\,\|(\lambda_3\mu_3)Y_3(w_3)I_3\rangle$$

$$= \frac{K(\lambda_1\mu_1)_{w_1 I_1} K(\lambda_2 0)_{w_2 I_2}}{K(\lambda_3\mu_3)_{w_3 I_3}} \sqrt{\frac{(w_1+w_2)!}{w_1! w_2!}} \begin{bmatrix} \frac{\lambda_1}{2} & \frac{w_1}{2} & I_1 \\ \frac{\lambda_2}{2} & \frac{w_2}{2} & I_2 \\ \frac{\lambda_3}{2} & \frac{w_1+w_2}{2} & I_3 \end{bmatrix} . \tag{73}$$

V. FORM II FOR THE SU(3)⊃SU(2)×U(1) WIGNER COEFFICIENT

The expression of Sec. IV is particularly simple if n, the number of squares added to row 3 of the Young tableau for $(\lambda_1\mu_1)$, is very small compared with $\lambda_2 - n$. In the case when $\lambda_2 - n \ll n$, it is advantageous to evaluate the Wigner coefficient

$$\langle (\lambda_3\mu_3) Y_3(w_3) I_3; (0\lambda_2) -\tfrac{1}{3}\lambda_2 + w_2, \quad I_2 = (\lambda_2 - w_2)/2 \| (\lambda_1\mu_1) Y_1(w_1) I_1 \rangle \tag{74}$$

via the reduced matrix element of the operator

$$T^{(0\lambda_2)}_{Y_2 = -(1/3)\lambda_2 + w_2, I_2 M_2} = \frac{(\alpha_{3a})^{w_2}}{\sqrt{w_2!}} P^{I_2 = \lambda_2/2 - w_2/2}_{M_2}(\alpha_a) . \tag{75}$$

The VCS realization of this operator is (again omitting the specific particle index, a)

$$\Gamma\left(\frac{(\alpha_3)}{\sqrt{w_2!}} P^{(\lambda_2 - w_2)/2}_{M_2}(\alpha) \right)$$
$$= \sum_{k=0}^{\lambda_2 - w_2} (-1)^k \sqrt{\frac{(\lambda_2 - w_2)!(w_2 + k)!}{(\lambda_2 - w_2 - k)! k! w_2!}} \frac{(\alpha_3)^{w_2 + k}}{\sqrt{(w_2+k)!}} [P^{(\lambda_2 - w_2 - k)/2}(\alpha) \times Z^{k/2}(z)]^{(\lambda_2 - w_2)/2}_{M_2} . \tag{76}$$

The needed intrinsic operator reduced matrix elements can be evaluated by the techniques illustrated in Sec. II. Now with $\lambda_1 + 2\mu_1 = \lambda_3 + 2\mu_3 + 2\lambda_2 - 3n'$,

$$\left((\lambda_1\mu_1) \frac{\lambda_1}{2} \left\| \frac{(\alpha_3)^{\lambda_2 - \nu}}{\sqrt{(\lambda_2 - \nu)!}} P^{\nu/2}(\alpha) \right\| (\lambda_3\mu_3) \left[\frac{n' - \nu}{2} \times \frac{\lambda_3}{2} \right] \frac{\lambda''}{2} \right)$$
$$= \left\langle (\lambda_1\mu_1) 0 \frac{\lambda_1}{2} \left\| \frac{\alpha_3^{\lambda_2 - n'}}{\sqrt{(\lambda_2 - n')!}} P^{n'/2}(\alpha) \right\| (\lambda_3\mu_3) 0 \frac{\lambda_3}{2} \right\rangle$$
$$\times \sqrt{\frac{(\lambda_2 - \nu)! n'!}{(\lambda_2 - n')!(n' - \nu)! \nu!}} \frac{K^2(\lambda_3, \mu_3 + \lambda_2 + 1 - \nu)_{n'(\lambda_1/2)}}{K^2(\lambda_3, \mu_3 + \lambda_2 + 1 - \nu)_{(n' - \nu)(\lambda''/2)} K^2(\lambda_3\mu_3)_{n'(\lambda_1/2)}} . \tag{77}$$

With these intrinsic operator matrix elements the recoupling techniques of Sec. IV can give the general Wigner coefficient (74). The symmetry property

$$\langle (\lambda_1\mu_1) Y_1(w_1) I_1; (\lambda_2 0)(\lambda_2/3) - w_2, I_2 = (\lambda_2 - w_2)/2 \| (\lambda_3\mu_3) Y_3(w_3) I_3 \rangle$$
$$= \sqrt{[\dim(\lambda_3\mu_3)/\dim(\lambda_1\mu_1)][(2I_1+1)/(2I_1+1)]} \, (-1)^{\lambda_1 + \mu_1 + \lambda_2 + \lambda_3 + \mu_3 + I_2 + I_3 - I_1}$$
$$\times \langle (\lambda_3\mu_3) Y_3(w_3) I_3; (0\lambda_2) - \lambda_2/3 + w_2, I_2 \| (\lambda_1\mu_1) Y_1(w_1) I_1 \rangle \tag{78}$$

gives the needed coefficient. Renaming $n' = \lambda_2 - n$ to be in agreement with the notation of Sec. IV we obtain the new form of the Wigner coefficient: with $\lambda_3 + 2\mu_3 = \lambda_1 + 2\mu_1 + \lambda_2 - 3n$, $w_3 = w_1 + w_2 - n$.

Form II:

$$\langle (\lambda_1\mu_1) Y_1(w_1) I_1; (\lambda_2 0)(\lambda_2/3) - w_2, I_2 = (\lambda_2 - w_2)/2 \| (\lambda_3\mu_3) Y_3(w_3) I_3) \rangle$$
$$= (-1)^{\mu_1 + \mu_3 + n + I_2 + I_3 - I_1} \sqrt{\frac{\dim(\lambda_3\mu_3)}{\dim(\lambda_1\mu_1)} \frac{2I_1 + 1}{2I_3 + 1}} \frac{K(\lambda_3\mu_3)_{w_3 I_3}}{K(\lambda_1\mu_1)_{w_1 I_1}}$$
$$\times [K(\lambda_3\mu_3)_{(\lambda_2 - n) I_3 = \lambda_1/2} K(\lambda_3, \mu_3 + 1 + n)_{\lambda_2 - n, I = \lambda_1/2}]^{-1}$$
$$\times \sum_{\nu=0}^{\min(\lambda_2 - n, \lambda_2 - w_2)} \sqrt{\frac{(\lambda_2 - w_2)!(\lambda_2 - n)!}{w_1! w_2! w_3! n!}} (-1)^\nu \frac{(\lambda_2 - \nu)!(w_1 + \lambda_2 - n - \nu)!}{\nu!(\lambda_2 - w_2 - \nu)!(\lambda_2 - n - \nu)!}$$

$$\times \sum_{\lambda''/2} \frac{K^2(\lambda_3,\mu_3+\lambda_2+1-\nu)_{(\lambda_2-n)(\lambda_1/2)}}{K^2(\lambda_3,\mu_3+\lambda_2+1-\nu)_{(\lambda_2-n-\nu)(\lambda''/2)}} U\left(\frac{\lambda_3}{2}\frac{\lambda_2-n-\nu}{2}\frac{\lambda_1}{2}\frac{\nu}{2};\frac{\lambda''}{2}\frac{\lambda_2-n}{2}\right)$$

$$\times(-1)^{\lambda_1/2-\lambda''/2-\lambda_3/2+n/2+\nu/2}\sqrt{\frac{(\lambda_1+1)(2I_2+1)(w_1+\lambda_2-n-\nu+1)}{(\lambda_3+1)(2I_2-\nu+1)(w_1+1)}}$$

$$\times\begin{bmatrix}\frac{\lambda_1}{2} & \frac{\nu}{2} & \frac{\lambda''}{2} & \frac{\lambda_3}{2}\\ I_1 & I_2 & I_3 & \frac{w_1+\lambda_2-n-\nu}{2}\\ \frac{w_1}{2} & I_2-\frac{\nu}{2} & \frac{\lambda_2-n-\nu}{2} & \frac{w_3}{2}\end{bmatrix}. \tag{79}$$

In the special case with $n=\lambda_2$ the sum in this expression collapses to a single term. The 12-j coefficient collapses to a 6-j coefficient, and with $(\lambda_1\mu_1)=(\lambda_3,\mu_3+n)$:

$$\langle(\lambda_3,\mu_3+n)Y_1(w_1)I_1;(n0)Y_2=n/3-w_2,I_2=(n-w_2)/2\|(\lambda_3\mu_3)Y_3(w_3)I_3\rangle$$

$$=\frac{K(\lambda_3\mu_3)_{w_3I_3}}{K(\lambda_3,\mu_3+n)_{w_1I_1}}\sqrt{\frac{(\lambda_3+\mu_3+2)(\mu_3+1)n!(w_1+1)!}{(\lambda_3+\mu_3+n+2)(\mu_3+n+1)(n-w_2)!w_2!(w_1+w_2-n+1)!}}$$

$$\times U\left(I_3\frac{n-w_2}{2}\frac{\lambda_3}{2}\frac{w_1}{2};I_1\frac{w_3}{2}\right) \tag{80}$$

This is the second special case of Le Blanc and Biedenharn.[11]

VI. FORM III. AN EXPRESSION WITH 9-j COEFFICIENTS ONLY

In the general case both forms I and II have the same complexity. Both involve two summations, the ν sum and a sum over an angular momentum quantum number. Both involve 12-j coefficients in the general case. Since 12-j coefficients may not be readily available a simpler expression would be useful. Such an expression can be derived by a buildup process in which the representation $(\lambda_2 0)$ is obtained from a stretched coupling of the representations $(n0)$ and $(\lambda_2-n,0)$, the first adding n squares to row 3 of the Young tableau for $(\lambda_1\mu_1)$ to make the representation (λ_1,μ_1-n), the second λ_2-n squares to rows 1 and 2 to make the final $(\lambda_3\mu_3)$:

$$\langle(\lambda_1\mu_1)Y_1(w_1)I_1;(\lambda_2 0)\tfrac{1}{3}\lambda_2-w_2,I_2=(\lambda_2-w_2)/2\|(\lambda_3\mu_3)Y_3(w_3)I_3\rangle U_{\mathrm{SU}_3}$$

$$=\sum_{w',I'}\langle(\lambda_1\mu_1)Y_1(w_1)I_1;(n0)\frac{n}{3}-w'',\frac{n-w''}{2}\|(\lambda_1,\mu_1-n)Y'(w')I'\rangle$$

$$\times\langle(\lambda_1,\mu_1-n)Y'(w')I';(\lambda_2-n,0)\frac{\lambda_2-n}{3}-w,\frac{\lambda_2-n-w}{2}\|(\lambda_3\mu_3)Y_3(w_3)I_3\rangle$$

$$\times\left\langle(n0)\frac{n}{3}-w''\frac{n-w''}{2};(\lambda_2-n,0)\frac{\lambda_2-n}{2}-w\frac{\lambda_2-n-w}{2}\middle|\middle|(\lambda_2 0)\frac{\lambda_2}{3}-w_2\frac{\lambda_2-w_2}{2}\right\rangle$$

$$\times U\left(I_1\frac{n-w''}{2}I_3\frac{\lambda_2-n-w}{2};I'\frac{\lambda_2-w_2}{2}\right), \tag{81}$$

with $w_3=w_1+w_2-n$, $w''=w'+n-w_1$, $w=w_3-w'$.

The coefficient for the coupling $(n0)\times(\lambda_2-n,0)\to(\lambda_2 0)$ is a special case of the 1st class of Le Blanc–Biedenharn. The needed 9-j coefficient is related by symmetry to a trivial 9-j coefficient with all stretched angular momentum couplings leading to the simple result (with $w_2=w+w''$)

$$\left\langle(n0)\frac{n}{3}-w'',\frac{n-w''}{2};(\lambda_2-n,0)\frac{\lambda_2-n}{3}-w,\frac{\lambda_2-n-w}{2}\middle|\middle|(\lambda_2 0)\frac{\lambda_2}{3}-w_2,\frac{\lambda_2-w_2}{2}\right\rangle$$

$$=\sqrt{\frac{(\lambda_2-n)!(\lambda_2-w_2)!n!w_2!}{(\lambda_2-n-w)!\lambda_2!(n-w'')!w!w''!}}. \tag{82}$$

Using Eqs. (73) and (80) for the remaining SU(3) Wigner coefficients, we obtain the right-hand side of Eq. (81).

The coefficient U_{SU_3} for the left-hand side of Eq. (81) can be treated as a normalization factor. It is the SU(3) U coefficient for the SU_3 recoupling implied by relation (81). It it were not known it could now be obtained from the right-hand

side of Eq. (81) for the special values $w_1 = 0, w_3 = 0, w_2 = n$ for which the 9-j coefficient and the U coefficients for the right-hand side are all unity. (In this case $w' = 0$ only and $I' = \lambda_1/2$ only.) From the known Wigner coefficient with $w_1 = w_3 = 0$ of Eq. (61), we can thus determine U_{SU_3} via this special case:

$$U_{SU_3}((\lambda_1\mu_1)(n0)(\lambda_3\mu_3)(\lambda_2 - n,0);(\lambda_1,\mu_1 - n)(\lambda_2 0))$$
$$= \sqrt{\frac{n!(\lambda_2 - n)!(\mu_1 - n + 1)!(\lambda_1 + \mu_1 + 2 - n)!}{\lambda_2!(\mu_1 + 1)!)(\lambda_1 + \mu_1 + 2)!} \frac{(\mu_3 + n + 1)!(\lambda_3 + \mu_3 + n + 2)!}{(\mu_3 + 1)!(\lambda_3 + \mu_3 + 2)!}}. \tag{83}$$

With this value we get the general result for the $SU(3) \supset SU(2) \times U(1)$ Wigner coefficient.

Form III (with $\lambda_3 + 2\mu_3 = \lambda_1 + 2\mu_1 + \lambda_2 - 3n$, $w_3 = w_1 + w_2 - n$):

$$\left\langle (\lambda_1\mu_1) Y_1(w_1) I_1; (\lambda_2 0) \frac{1}{3}\lambda_2 - w_2, I_2 = \frac{\lambda_2 - w_2}{2} \middle|\middle| (\lambda_3\mu_3) I_3 \right\rangle$$
$$= \sqrt{\frac{(\lambda_1 + \mu_1 + 1)!\mu_1!(\mu_3 + 1)!(\lambda_3 + \mu_3 + 2)!(\lambda_2 - n + 1)!(\lambda_2 - w_2)!}{(\lambda_1 + \mu_1 + 1 - n)!(\mu_1 - n)!(\mu_3 + n + 1)!(\lambda_3 + \mu_3 + n + 2)!}}$$
$$\times \sum_{w'} \sum_{I'} \sqrt{\frac{n!(w_1 + 1)!w_3!w_2!}{(\lambda_2 - n - w_3 + w' + 1)(w' + 1)}} \frac{1}{w'!(\lambda_2 - n - w_3 + w')!(w_1 - w')!(n + w' - w_1)!(w_3 - w')!}$$
$$\times \frac{K^2(\lambda_1,\mu_1 - n)_{w'I'}}{K(\lambda_1\mu_1)_{w_1 I_1} K(\lambda_3\mu_3)_{w_3 I_3}} \begin{bmatrix} \frac{\lambda_1}{2} & \frac{w'}{2} & I' \\ \frac{\lambda_2 - n}{2} & \frac{w_3 - w'}{2} & \frac{\lambda_2 - n - w_3 + w'}{2} \\ \frac{\lambda_3}{2} & \frac{w_3}{2} & I_3 \end{bmatrix}$$
$$\times U\left(I_1 \frac{w_1 - w'}{2} I_3 \frac{\lambda_2 - n - w_3 + w'}{2}; I' \frac{\lambda_2 - w_2}{2}\right) U\left(I' \frac{w_1 - w'}{2} \frac{\lambda_1}{2} \frac{w_1}{2}; I_1 \frac{w'}{2}\right). \tag{84}$$

For the special case with $w_3 = 0$ this collapses to the simple result:

$$\left\langle (\lambda_1\mu_1) Y_1(w_1) I_1; (\lambda_2 0) \frac{1}{3}\lambda_2 - w_2, I_2 = \frac{\lambda_2 - w_2}{2} \middle|\middle| (\lambda_3\mu_3) \frac{1}{3}(\lambda_3 + 2\mu_3) I_3 = \frac{\lambda_3}{2} \right\rangle$$
$$= (-1)^{\lambda_1/2 + w_1/2 - I_1} \sqrt{\frac{(2I_1 + 1)n! - (\lambda_2 - w_2)!}{(\lambda_1 + 1)(\lambda_2 - n)!w_1!(n - w_1)!}}$$
$$\times \sqrt{\frac{(\lambda_1 + \mu_1 + 1)!\mu_1!(\lambda_3 + \mu_3 + 2)!(\mu_3 + 1)!}{(\lambda_1 + \mu_1 + 1 - n)!(\mu_1 - n)!(\lambda_3 + \mu_3 + n + 2)!(\mu_3 + n + 1)!}}$$
$$\times \frac{1}{K(\lambda_1\mu_1)_{w_1 I_1}} U\left(I_1 \frac{w_1}{2} \frac{\lambda_3}{2} \frac{\lambda_2 - n}{2}; \frac{\lambda_1}{2} \frac{\lambda_2 - w_2}{2}\right). \tag{85}$$

For the special case with $w_1 = 0$, on the other hand, Eq. (84) leads to

$$\left\langle (\lambda_1\mu_1) \frac{1}{3}(\lambda_1 + 2\mu_1), I_1 = \frac{\lambda_1}{2}; (\lambda_2 0) \frac{\lambda_2}{3} - w_2, I_2 = \frac{\lambda_2 - w_2}{2} \middle|\middle| (\lambda_3\mu_3) Y_3(w_3) I_3 \right\rangle$$
$$= \sqrt{\frac{(\lambda_2 - n + 1)!w_2!(\lambda_1 + \mu_1 + 1)!\mu_1!(\lambda_3 + \mu_3 + 2)!(\mu_3 + 1)!}{(\lambda_2 - w_2 + 1)!n!(\lambda_1 + \mu_1 + 1 - n)!(\mu_1 - n)!(\mu_3 + n + 1)!(\lambda_3 + \mu_3 + n + 2)!w_3!}}$$
$$\times \frac{1}{K(\lambda_3\mu_3)_{w_3 I_3}} U\left(\frac{\lambda_1}{2} \frac{\lambda_2 - n}{2} I_3 \frac{w_3}{2}; \frac{\lambda_3}{2} \frac{\lambda_2 - w_2}{2}\right). \tag{86}$$

Finally, by interchanging the order of the coupling to $(\lambda_2 - n,0) \times (n0)$ in the analog of Eq. (81) still another form can be obtained for the totally symmetric $SU(3) \supset SU(2) \times U(1)$ Wigner coefficient.

Form III' (with $\lambda_3 + 2\mu_3 = \lambda_1 + 2\mu_1 + \lambda_2 - 3n$, $w_3 = w_1 + w_2 - n$):

$$\left\langle (\lambda_1\mu_1)Y_1(w_1)I_1;(\lambda_2 0)\,\frac{1}{3}\lambda_2 - w_2, I_2 = \frac{\lambda_2 - w_2}{2}\middle|\middle|\,(\lambda_3\mu_3)Y_3(w_3)I_3\right\rangle$$

$$= \sqrt{\frac{(\lambda_3+\mu_3+2)(\mu_3+1)(\lambda_3+\mu_3+n+1)!(\mu_3+n)!(\mu_1-n)!(\lambda_1+\mu_1+1-n)!}{(\lambda_3+\mu_3+n+2)(\mu_3+n+1)(\lambda_3+\mu_3+1)!\mu_3!\mu_1!(\lambda_1+\mu_1+1)!}}$$

$$\times\sqrt{\frac{(\lambda_2-n+1)!(\lambda_2-w_2)!n!w_2!}{w_1!(w_3+1)!}}\sum_{w,I}\frac{(w_1+w)!}{w!(w_2-w)!(n-w_2+w)!(\lambda_2-n-w)!}$$

$$\times\sqrt{\frac{(w_1+w+1)}{(\lambda_2-n+1-w)}}\times\frac{K(\lambda_1\mu_1)_{w_1,I_1}K(\lambda_3\mu_3)_{w_3,I_3}}{K^2(\lambda_3,\mu_3+n)_{w+w_1,I}}\begin{bmatrix}\frac{\lambda_1}{2} & \frac{w_1}{2} & I_1\\ \frac{\lambda_2-n}{2} & \frac{w}{2} & \frac{\lambda_2-n-w}{2}\\ \frac{\lambda_3}{2} & \frac{w_1+w}{2} & I\end{bmatrix}$$

$$\times U\left(I_1\,\frac{\lambda_2-n-w}{2}\,I_3\,\frac{n-w_2+w}{2};I\,\frac{\lambda_2-w_2}{2}\right)U\left(I_3\,\frac{n-w_2+w}{2}\,\frac{\lambda_3}{2}\,\frac{w_1+w}{2};I\,\frac{w_3}{2}\right). \tag{87}$$

VII. SUMMARY

Three types of expressions have been derived within the framework of generalized VCS theory for the SU(3) $\supset$SU(2)$\times$U(1) Wigner coefficients for the multiplicity-free coupling $(\lambda_1\mu_1)\times(\lambda_2 0)\to(\lambda_3\mu_3)$ involving totally symmetric U(3) tensors. All three involve two summations and are therefore comparable to previously known results[15] as far as their complexity is concerned. All results are expressed in terms of SU(2) recoupling coefficients and the simple K-normalization factors of VCS theory and therefore throw new light on the structure of such coefficients. Two of the expressions, given by Eqs. (72) and (79), involve 12-j coefficients. Their main value lies in the fact that they illustrate how the spectacularly simple special cases of Le Blanc and Biedenharn[11] arise as special cases of very general results. Simpler expressions involving only 9-j coefficients and Racah coefficients of 6-j type are given by Eqs. (84) and (87). These are derived by a coupling process which compounds the two special Le Blanc–Biedenharn results. In this process a U(3) tensor that can add squares only to row 3 of the starting tableau is combined with a U(3) tensor that can add squares only to rows 1 and 2. Since tensors for the generic case with multiplicity[14] can also be built in this fashion the VCS techniques used in this investigation may be useful for the general coupling with multiplicity. Special values for the SU(3) Wigner coefficients for the coupling $(\lambda_1\mu_1)\times(\lambda_2 0)\to(\lambda_3\mu_3)$ in which either the $(\lambda_1\mu_1)$ or $(\lambda_3\mu_3)$ states are restricted to highest weight have also been given in a new form involving a simple Racah coefficient, see Eqs. (85) and (86).

Note added in proof: The Wigner coefficients of this investigation use the phase convention of Draayer and Akiyama (see Ref. 5). To convert to the phase convention of Biedenharn and Louck (Refs. 8–10), the state vector of Eq. (24) must be multiplied by the phase $(-1)^{\lambda/2-\omega/2-I}$ leading to an additional overall phase of $(-1)^{\lambda_1/2-\omega_1/2-I_1-\lambda_3/2+\omega_3/2+I_3}$ to convert the Wigner coefficients of this investigation to the Biedenharn–Louck convention.

ACKNOWLEDGMENTS

K. T. Hecht was supported in part by the U.S. National Science Foundation and L. C. Biedenharn was supported in part by the Department of Energy.

[1] D. J. Rowe, J. Math. Phys. **25**, 2662 (1984).
[2] D. J. Rowe, G. Rosensteel, and R. Gilmore, J. Math. Phys. **26**, 2787 (1985).
[3] J. Deenen and C. Quesne, J. Math. Phys. **25**, 1638, 2354 (1984).
[4] C. Quesne, J. Math. Phys. **27**, 869 (1986).
[5] K. T. Hecht, *The Vector Coherent State Method and its Application to Problems of Higher Symmetries*, Lecture Notes in Physics, Vol. 290 (Springer, Berlin, 1987).
[6] D. J. Rowe, R. Le Blanc, and K. T. Hecht, J. Math. Phys. **29**, 287 (1988).
[7] A. Klein, T. D. Cohen, and C. T. Li, Ann. Phys. **141**, 382 (1982).
[8] K. T. Hecht, R. Le Blanc, and D. J. Rowe, J. Phys. A: Math. Gen. **20**, 2241 (1987).
[9] R. Le Blanc and K. T. Hecht, J. Phys. A: Math. Gen. **20**, 4613 (1987).
[10] L. C. Biedenharn and J. D. Louck, Commun. Math. Phys. **8**, 89 (1972).
[11] R. Le Blanc and L. C. Biedenharn, J. Phys. A: Math. Gen. **22**, 31 (1989).
[12] K. T. Hecht, Nucl. Phys. A **493**, 29 (1989).
[13] K. T. Hecht and J. Q. Chen, Nucl. Phys. A **512**, 365 (1990).
[14] L. C. Biedenharn, R. Le Blanc, and J. D. Louck, "Recent Progress in Implementing the Tensor Operator Calculus," in Proc. of Symmetries in Science III, Bregenz, Austria (1988).
[15] For a good summary, see: Y. Fujiwara and H. Horiuchi, "Properties of Double Gel'fand Polynomials and their Applications to Multiplicity-free Problems," Mem. Faculty Sci., Kyoto Univ. A **36** (2), Article 1 (1983).
[16] L. C. Biedenharn, A. Giovannini, and J. D. Louck, J. Math. Phys. **8**, 691 (1967).
[17] R. Le Blanc and D. J. Rowe, J. Phys. A: Math. Gen. **19**, 1093, 1111 (1986).
[18] J. P. Elliot, Proc. R. Soc. London Ser. A **245**, 128, 562 (1958).
[19] H. A. Jahn and J. Hope, Phys. Rev. **93**, 318 (1954).
[20] R. J. Ord-Smith, Phys. Rev. **94**, 1227 (1954).

New inhomogeneous boson realizations and inhomogeneous differential realizations of Lie algebras

Hong-Chen Fu and Chang-Pu Sun
CCAST (World Laboratory), P. O. Box 8730, Beijing, People's Republic of China, and Department of Physics, Northeast Normal University, Changchun, Jilin Province 130024, People's Republic of China[a)]

(Received 22 September 1989; accepted for publication 11 July 1990)

The inhomogeneous boson realizations (IHBR) and the corresponding inhomogeneous differential realizations (IHDR) of Lie algebras, which play an important role in the search of quasi-exactly solvable problems (QESP) of quantum mechanics, are studied. All possible IHDR of semisimple Lie algebras can be obtained in this way. As examples, the IHBR and the corresponding IHDR of Lie algebras SU(2) and SU(3) are studied in detail.

I. INTRODUCTION

Recently discovered quasi-exactly solvable problems (QESP)[1-4] of quantum mechanics have been proved to be related to the inhomogeneous differential realizations (IHDR) of Lie algebras.[2-4] Turbiner studied the one-dimensional QESP by making use of the IHDR of Lie algebra sl(2),[2] and pointed out that one could study the multi-dimensional QESP by using the IHDR of sl(m) algebra.[2] Shifman and Turbiner studied the two-dimensional QESP by making use of the IHDR of Lie algebras su(2)$\times$su(2), so(3), and su(3).[3,4] Construction of IHDR of Lie algebras is very important in the search of QESP. By extending Shifman's discussions, the authors of this paper have obtained the IHDR of any Lie algebras.[5] However, as pointed in Ref. 5, this IHDR is trivial, i.e., it does not include all possible representations of Lie algebras. In this paper, we will continue studying the IHDR of Lie algebras.

On the universal enveloping algebra $\mathscr{U}(L)$ of Lie algebra L, or on its quotient space $\mathscr{U}(L)/J$, where J is a left ideal generated through algebraic relations in $\mathscr{U}(L)$, one can obtain all the representations of L, including indecomposable and irreducible representations. These representations are well-defined. Especially, for some concrete Lie algebras, Gruber *et al.* have given the explicit expressions of these representations.[6-9] In this paper, from a representation on $\mathscr{U}(L)$ or on $\mathscr{U}(L)/J$, we define a representation of Lie algebra L on the Fock space, which is automorphic to $\mathscr{U}(L)$ or $\mathscr{U}(L)/J$, then give the IHBR of Lie algebra L. By making use of the corresponding relation between creation and annihilation operators of boson states and differential operators, we obtain the IHDR of Lie algebras.

For semisimple Lie algebras, further discussions are given. On the Fock space that corresponds to the finite-dimensional irreducible standard cyclic module $V(\lambda)$[10] as the quotient space of $\mathscr{U}(L)$, all possible IHBR of semisimple Lie algebras marked by rank L non-negative integers, where rank L is the rank of semisimple Lie algebra L, are obtained. The IHDR of L obtained from the corresponding IHBR will use $\frac{1}{2}$(dim L-rank L) independent variables, which is in accord with Shifman's inference.

As examples, we discuss varieties of IHBR and IHDR of Lie algebras SU(2) and SU(3) by making use of their representations on their universal enveloping algebras or on their quotient spaces given by Gruber *et al.*[6,7] The IHDR of SU(3) marked by two non-negative integers can be obtained.

It is worth noticing that Doebner *et al.* have studied the IHBR and IHDR of SU(2) and SU(1,1) from the representations of SU(2) and SU(1,1) on their universal enveloping algebras or on their quotient spaces.[11] Their results are the Hermitian conjugate of our results, combined with the index change. However, their method is different from the method used in this paper. One will find that our method is straightforward and simpler than that used in Ref. 11.

This paper is organized as follows. After studying in Sec. II the general procedure to construct the IHBR and IHDR of any Lie algebra, we will further show this procedure with SU(2) as an example in Sec. III. In Sec. IV, the further discussions for the semisimple Lie algebras are given. In Sec. V, the IHBR and IHDR of Lie algebra SU(3) are studied in detail.

The symbol $\mathbb{Z}^+$ denotes the set of non-negative integers. The symbol $\mathbb{C}$ denotes the complex number field.

II. GENERAL PROCEDURE

Let the basis for Lie algebra L be $\{T_a | a = 1,2,...,M, \dim L = M\}$. According to PBW theorem, the basis for the universal enveloping algebra $\mathscr{U}(L)$ of Lie algebra L can be chosen as

$$\{X(i_1,i_2,...,i_M) \equiv T_1^{i_1} T_2^{i_2} \cdots T_M^{i_M} | i_1,i_2,...,i_M \in \mathbb{Z}^+\}. \tag{2.1}$$

If we regard $\mathscr{U}(L)$ as the left L module, namely,

$$\begin{aligned}\rho(T_a) X(i_1,i_2,...,i_M) &= T_a T_1^{i_1} T_2^{i_2} \cdots T_M^{i_M} \\ &= \sum_{i_1' \cdots i_M'} \rho(T_a)_{i_1 i_2 \cdots i_M}^{i_1' i_2' \cdots i_M'} X(i_1',i_2',...,i_M'),\end{aligned} \tag{2.2}$$

[a)] Mailing address.

 0022-2488/90/122797-06$03.00

one can obtain a representation ρ of L on $\mathscr{U}(L)$, called master representation. The matrix elements $\rho(T_a)_{i_1 i_2 \cdots i_M}^{i'_1 i'_2 \cdots i'_M}$ are determined by the Lie product of L and are related to the M non-negative integers $i_1, i_2, \ldots, i_M$.

Let J be a left ideal of $\mathscr{U}(L)$ generated through algebraic relations in $\mathscr{U}(L)$. On the quotient space $\mathscr{U}(L)/J$, which basis is also marked by certain of non-negative integers, (2.2) induces a representation. The different choice of J enables us to obtain all possible representations of L. For some concrete Lie algebras, these representations have been written as the explicit expressions.[6-9] Now we construct the IHBR of L from these representations. However, for the sake of convention we only construct the IHBR of L from the master representation even though the arguments work also for the representation on $\mathscr{U}(L)/J$.

We first define a Fock space $\mathscr{F}$ with basis

$$\{|i_1,i_2,\ldots,i_M\rangle \equiv a_1^{+i_1} a_2^{+i_2} \cdots a_M^{+i_M}|0\rangle \,|a_i|0\rangle = 0,\ i_k \in \mathbb{Z}^+\}. \tag{2.3}$$

Then the mapping φ: $\mathscr{U}(L) \to \mathscr{F}$ defined by

$$\varphi(X(i_1,i_2,\ldots,i_M)) = |i_1,i_2,\ldots,i_M\rangle \tag{2.4}$$

is an associative algebraic isomorphism. Let

$$\Gamma(T_a) = \varphi\rho(T_a)\varphi^{-1}. \tag{2.5}$$

Then $\Gamma(T_a)$ is a linear operator on $\mathscr{F}$ and satisfies the relation

$$[\Gamma(T_a),\Gamma(T_b)] = \Gamma([T_a,T_b]). \tag{2.6}$$

Therefore Eq. (2.5) defines a representation of L on $\mathscr{F}$, called Fock representation. It is easy to see that

$$\Gamma(T_a)|i_1,i_2,\ldots,i_M\rangle = \sum_{i'_1 i'_2 \cdots i'_M} \rho(T_a)_{i_1 i_2 \cdots i_M}^{i'_1 i'_2 \cdots i'_M} |i'_1,i'_2,\ldots,i'_M\rangle. \tag{2.7}$$

From (2.7) it follows that

$$\begin{aligned}
\Gamma(T_a) &= \sum_{\substack{i'_1 i'_2 \cdots i'_M \\ j_1 j_2 \cdots j_M}} \rho(T_a)_{j_1 j_2 \cdots j_M}^{i'_1 i'_2 \cdots i'_M} \frac{1}{j_1! j_2! \cdots j_M!} |i'_1,\ldots,i'_M\rangle\langle j_1 j_2,\ldots j_M| \\
&= \sum_{\substack{i'_1 i'_2 \cdots i'_M \\ j_1 j_2 \cdots j_M}} \rho(T_a)_{j_1 j_2 \cdots j_M}^{i'_1 i'_2 \cdots i'_M} \frac{1}{j_1! j_2! \cdots j_M!} a_1^{+i'_1} a_2^{+i'_2} \cdots a_M^{+i'_M} |0,0,\ldots,0\rangle\langle 0,0,\ldots,0| a_1^{j_1} a_2^{j_2} \cdots a_M^{j_M} \\
&= \sum_{\substack{i'_1 i'_2 \cdots i'_M \\ j_1 j_2 \cdots j_M}} \rho(T_a)_{j_1 j_2 \cdots j_M}^{i'_1 i'_2 \cdots i'_M} \frac{1}{j_1! j_2! \cdots j_M!} a_1^{+i'_1} a_2^{+i'_2} \cdots a_M^{+i'_M} :\exp\Big(-\sum_{i=1}^{M} a_i^+ a_i\Big): a_1^{j_1} a_2^{j_2} \cdots a_M^{j_M},
\end{aligned} \tag{2.8}$$

where the formula

$$|0,0,\ldots,0\rangle\langle 0,0,\ldots,0| = :\exp\Big(-\sum_{i=1}^{M} a_i^+ a_i\Big): \tag{2.9}$$

is vacuum projective operator, $:\cdots:$ is normal product. Equation (2.8) is the required IHBR of L.

However, for a concrete Lie algebra, it is inconvenient to achieve the IHBR by use of Eq. (2.8). In fact, when we know the explicit form of $\rho(T_a)_{j_1 \cdots j_M}^{i_1 \cdots i_M}$, we can easily obtain the IHBR of L. In this case the following formulas usually used:

$$\begin{aligned}
&a_k^+ |i_1,i_2,\ldots,i_M\rangle = |i_1,i_2,\ldots,i_k+1,\ldots,1_M\rangle, \\
&a_k |i_1,i_2,\ldots,i_M\rangle = i_k |i_1,\ldots,i_k-1,\ldots,i_M\rangle, \\
&a_k^+ a_k |i_1,i_2,\ldots,i_M\rangle = i_k |i_1,i_2,\ldots,i_M\rangle, \\
&a_k a_k^+ |i_1,i_2,\ldots,i_M\rangle = (i_k+1)|i_1,i_2,\ldots,i_M\rangle, \\
&e^{a_k^+} |i_1,i_2,\ldots,i_M\rangle \\
&\quad = \sum_{j=0}^{\infty} \frac{1}{j!} |i_1,i_2,\ldots,i_k+j,\ldots,i_M\rangle, \\
&e^{ak} |i_1,i_2,\ldots,i_M\rangle \\
&\quad = \sum_{j=0}^{i_k} \frac{i_k!}{j!(i_k-j)!} |i_1,i_2,\ldots,j,\ldots,i_M\rangle.
\end{aligned} \tag{2.10}$$

In the next section this technique will be shown by virtue of SU(2) as an example, and in Sec. V with SU(3) as an example.

From the IHBR of L, we can immediately write the IHDR of L by making use of the following corresponding relations:

$$\xi_i \Leftrightarrow a_i^+, \quad \frac{\partial}{\partial \xi_i} \Leftrightarrow a_i \quad (i = 1,2,\ldots,M). \tag{2.11}$$

III. IHBR AND IHDR OF LIE ALGEBRA SU(2)

Let the basis for Lie algebra SU(2) be $\{T_+, T_-, T_0\}$ with the following commutation relations:

$$[T_0,T_\pm] = \pm 2T_\pm, \quad [T_+,T_-] = T_0. \tag{3.1}$$

The basis for the universal enveloping algebra $\mathscr{U}$(SU(2)) of SU(2) can be chosen as

$$\{X(m,n,r) \equiv T_+^m T_-^n T_0^r | m,n,r \in \mathbb{Z}^+\}. \tag{3.2}$$

(1) From the master representation of SU(2) on $\mathscr{U}$(SU(2)) given in Ref. 6

$$\begin{aligned}
&\rho(T_0)X(m,n,r) = X(m,n,r+1) + 2(m-n)X(m,n,r), \\
&\rho(T_+)X(m,n,r) = X(m+1,n,r), \\
&\rho(T_-)X(m,n,r) = X(m,n+1,r) - mX(m-1,n,r+1) \\
&\qquad\qquad + m(2n-m+1)X(m-1,n,r),
\end{aligned} \tag{3.3}$$

the corresponding Fock representation is obtained as

$$\Gamma(T_0)|m,n,r\rangle = |m,n,r+1\rangle + 2(m-n)|m,n,r\rangle,$$
$$\Gamma(T_+)|m,n,r\rangle = |m+1,n,r\rangle,$$
$$\Gamma(T_-)|m,n,r\rangle = |m,n+1,r\rangle - m|m-1,n,r+1\rangle + m(2n-m+1)|m-1,n,r\rangle. \tag{3.4}$$

From (3.4) we can immediately obtain the IHBR as

$$\Gamma(T_0) = a_3^+ + 2a_1^+ a_1 - 2a_2^+ a_2,$$
$$\Gamma(T_+) = a_1^+, \tag{3.5}$$
$$\Gamma(T_-) = a_2^+ - a_3^+ a_1 + 2a_2^+ a_2 a_1 - 2a_1^+ a_2^2.$$

The corresponding IHDR is obtained as

$$D(T_0) = \xi_3 + 2\xi_1\frac{\partial}{\partial\xi_1} - 2\xi_2\frac{\partial}{\partial\xi_2},$$
$$D(T_+) = \xi_1, \tag{3.6}$$
$$D(T_-) = \xi_2 - \xi_3\frac{\partial}{\partial\xi_1} + 2\xi_2\frac{\partial^2}{\partial\xi_2\,\partial\xi_1} - 2\xi_1\frac{\partial^2}{\partial\xi_1^2}.$$

In comparison with the result in Ref. 11, we will find that (3.5) and (3.6) are the Hermitian conjugate of Doebner's result, combined with an index change $1\leftrightarrow 2$.

(2) From the representation[6] of SU(2) on the quotient space $\mathcal{U}(\mathrm{SU}(2))/I_1$, where I_1 is a left ideal of $\mathcal{U}(\mathrm{SU}(2))$ generated by $T_0 - \Lambda\cdot 1$ ($\Lambda\in\mathbb{C}$), with basis

$$\{X(m,n)\equiv X(m,n,0)\,\mathrm{Mod}\,\mathrm{I}\,|\,m,n\in\mathbb{Z}^+\}, \tag{3.7}$$
$$\rho(T_0)X(m,n) = [\Lambda + 2(m-n)]X(m,n),$$
$$\rho(T_+)X(m,n) = X(m+1,n),$$
$$\rho(T_-)X(m,n) = X(m,n+1) + m(-\Lambda + 2n - m + 1)\times X(m-1,n), \tag{3.8}$$

we can obtain the Fock representation. From this Fock representation the IHBR of SU(2) is obtained as

$$\Gamma(T_0) = \Lambda + 2a_1^+ a_1 - 2a_2^+ a_2,$$
$$\Gamma(T_+) = a_1^+, \tag{3.9}$$
$$\Gamma(T_-) = a_2^+ - \Lambda a_1 + 2a_2^+ a_2 a_1 - a_1^+ a_1^2.$$

The corresponding IHDR is

$$D(T_0) = \Lambda + 2\xi_1\frac{\partial}{\partial\xi_1} - 2\xi_2\frac{\partial}{\partial\xi_2},$$
$$D(T_+) = \xi_1, \tag{3.10}$$
$$D(T_-) = \xi_2 - \Lambda\frac{\partial}{\partial\xi_1} + 2\xi_2\frac{\partial^2}{\partial\xi_2\,\partial\xi_1} - \xi_1\frac{\partial^2}{\partial\xi_1^2}.$$

(3) The master representation ρ induces on the quotient space $\mathcal{U}(\mathrm{SU}(2))/I_2$, where I_2 is a left ideal generated by $T_- - \lambda 1$ ($\lambda\in\mathbb{C}$), with basis

$$\{X(m,r)\equiv X(m,0,r)\,\mathrm{Mod}\,I_2\,|\,m,r\in\mathbb{Z}^+\} \tag{3.11}$$

a representation[6]

$$\rho(T_0)X(m,r) = X(m,r+1) + 2mX(m,r),$$
$$\rho(T_+)X(m,r) = X(m+1,r),$$
$$\rho(T_0)X(m,r) = \lambda\sum_{k=0}^{r}\frac{r!}{(r-k)!k!}X(m,k) - mX(m-1,r+1) - m(m-1)X(m-1,r). \tag{3.12}$$

From the Fock representation that corresponds to (3.12) one obtains the IHBR of SU(2) as

$$\Gamma(T_0) = a_2^+ + 2a_1^+ a_1,$$
$$\Gamma(T_+) = a_1^+, \tag{3.13}$$
$$\Gamma(T_-) = \lambda e^{a_2^+} - a_2^+ a_1 - a_1^+ a_1^2,$$

The corresponding IHDR of SU(2) is obtained as

$$D(T_0) = \xi_2 + 2\xi_1\frac{\partial}{\partial\xi_1},$$
$$D(T_+) = \xi_1, \tag{3.14}$$
$$D(T_-) = \lambda e^{\partial/\partial\xi_2} - \xi_2\frac{\partial}{\partial\xi_1} - \xi_1\frac{\partial^2}{\partial\xi_1^2},$$

(4) Let I_3 be a left ideal generated by T_-, $T_0 - \Lambda$ ($\Lambda\in\mathbb{C}$). Then on the quotient space $V\equiv\mathcal{U}(\mathrm{SU}(2))/I_3$ with basis

$$\{X(m)\equiv X(m,0,0)\,\mathrm{Mod}\,I_3\,|\,m\in\mathbb{Z}^+\}, \tag{3.15}$$

the master representation ρ induces a representation[6]

$$\rho(T_0)X(m) = (\Lambda + 2m)X(m),$$
$$\rho(T_+)X(m) = X(m+1), \tag{3.16}$$
$$\rho(T_-)X(m) = m(-\Lambda - m + 1)X(m-1).$$

From corresponding Fock representation we obtain the IHBR and IHDR as

$$\Gamma(T_0) = \Lambda + 2a^+ a, \quad D(T_0) = \Lambda + 2\xi\frac{\partial}{\partial\xi},$$
$$\Gamma(T_+) = a^+, \quad D(T_+) = \xi, \tag{3.17}$$
$$\Gamma(T_-) = -\Lambda a - a^+ a^2, \quad D(T_-) = -\Lambda\frac{\partial}{\partial\xi} - \xi\frac{\partial^2}{\partial\xi^2}.$$

(5) When $\Lambda = -N$, where $N\in\mathbb{Z}^+$, the subspace V_N of V spanned by

$$V_N:\{X(m)\in V\,|\,m\geqslant N+1, m\in\mathbb{Z}^+\} \tag{3.18}$$

is invariant under the action of (3.16). On the quotient space $\tilde{V}_N\equiv V/V_N$ with basis

$$\tilde{V}_N:\{X_N(m)\equiv(N-m)!X(m)\,\mathrm{Mod}\,V_N\,|\,m=0,1,2,\ldots,N\},$$
$$\dim\tilde{V}_N = N+1, \tag{3.19}$$

(3.16) subduces a finite-dimensional representation[6]

$$\rho(T_0)X_N(m) = (-N/2+m)2X_N(m),$$
$$\rho(T_+)X_N(m) = (N-m)X_N(m+1), \tag{3.20}$$
$$\rho(T_-)X_N(m) = mX_N(m-1).$$

The corresponding Fock representation is obtained as

$$\Gamma(T_0')|m\rangle = (-N/2+m)|m\rangle,$$
$$\Gamma(T_+)|m\rangle = (N-m)|m+1\rangle, \tag{3.21}$$
$$\Gamma(T_-)|m\rangle = m|m-1\rangle \quad (T_0' = \tfrac{1}{2}T_0).$$

It should be noted that the space carrying the representation (3.21) is a $(N+1)$-dimensional invariant subspace $\mathscr{F}_N$ spanned by $\{|m\rangle | m=0,1,...,N\}$ of Fock space with basis

$$\mathscr{F}:\{|m\rangle = a^{+m}|0\rangle | m\in\mathbb{Z}^+,\ a|0\rangle = 0\}. \tag{3.22}$$

From the finite-dimensional Fock representation (3.21) we obtain the finite-dimensional IHBR and IHDR as

$$\Gamma(T_0') = -N/2 + a^+a, \quad D(T_0') = -\frac{N}{2} + \xi\frac{\partial}{\partial\xi},$$

$$\Gamma(T_+) = Na^+ - a^{+2}a, \quad D(T_+) = N\xi - \xi^2\frac{\partial}{\partial\xi},$$

$$\Gamma(T_-) = a, \quad D(T_-) = \frac{\partial}{\partial\xi}, \tag{3.23}$$

which are just the result obtained in Ref. 5.

IV. COMMENTS ON IHDR OF SEMISIMPLE LIE ALGEBRAS

Let L be a semisimple Lie algebra with Cartan decomposition $L = H \oplus \Sigma_{\beta\in\Phi} L_\beta$, where H is Cartan subspace ($\dim H = l = \text{rank } L$), Φ is root system, L_β is the root space of root β. Let $\Phi^+ = \{\beta_1,\beta_2,...,\beta_m\}$ be a set of positive roots and $\Delta = \{\alpha_1,\alpha_2,...,\alpha_l\}\subset\Phi$ the simple root system. Choose the basis for L as

$$\{y_{\beta_1},y_{\beta_2},...,y_{\beta_m},h_1,h_2,...,h_l,x_{\beta_1},x_{\beta_2},...,x_{\beta_m}\}, \tag{4.1}$$

where $y_{\beta_i}\in L_{-\beta_i}$, $x_{\beta_i}\in L_{\beta_i}$, $h_i\equiv h_{\alpha_i}\in H$, with the following Lie product:

$$[h_i,h_j] = 0 \quad (i,j = 1,2,...,l), \quad [h_i,x_{\beta_j}] = \beta_j(h_i)x_{\beta_j},$$
$$[h_i,y_{\beta_j}] = -\beta_j(h_i)y_{\beta_j}, \quad [x_{\beta_i},y_{\beta_j}] = \delta_{ij}h_{\beta_j}. \tag{4.2}$$

If we regard the universal enveloping algebra $\mathscr{U}(L)$ of L with PBW basis

$$\{y_{\beta_1}^{i_1}y_{\beta_2}^{i_2}\cdots y_{\beta_m}^{i_m}h_1^{k_1}h_2^{k_2}\cdots h_l^{k_l}x_{\beta_1}^{j_1}x_{\beta_2}^{j_2}\cdots x_{\beta_m}^{j_m} | i_p,k_p,j_p\in\mathbb{Z}^+\} \tag{4.3}$$

as the left L module, we obtain the master representation of L.

Let $I(\lambda)$ ($\lambda\in H^*$) be a left ideal generated by $\{x_{\beta_i},h_\alpha - \lambda(h_\alpha)1 | \beta_i\in\Phi^+,\alpha\in\Phi\}$. Then the quotient space $Z(\lambda) = \mathscr{U}(L)/I(\lambda)$ with basis

$$\{X(i_1,i_2,...,i_m)\equiv[y_{\beta_1}^{i_1},y_{\beta_2}^{i_2}\cdots y_m^{i_m}]\text{Mod } I(\lambda) | ik\in\mathbb{Z}^+\} \tag{4.4}$$

is a indecomposable standard cyclic module with the highest weight λ.[10]

If λ is a dominant integral linear function, i.e., $\lambda(h_i)\equiv\Lambda_i\in\mathbb{Z}^+$, the left ideal $Y(\lambda)$ generated by $\{y_{\alpha_c}^{\Lambda_i+1}\text{Mod}I(\lambda) | i = 1,2,...,l\}$ is the unique maximal proper submodule and the quotient module $V(\lambda)\equiv Z(\lambda)/Y(\lambda)$ spanned by

$$\{B(i_1,i_2,...,i_m) \equiv X(i_1,i_2,...,i_m)\text{Mod } Y(\lambda) | i_1,i_2,...,i_m\in\mathbb{Z}^+\} \tag{4.5}$$

is a finite-dimensional irreducible L module,[10] on which the representation is marked by $l = \text{rank } L$ non-negative integers $(\Lambda_1,\Lambda_2,...,\Lambda_l)$.

From the representation on $V(\lambda)$, we can obtain the IHBR and the corresponding IHDR of L by use of the procedure given in Sec. II, which are marked by $(\Lambda_1,\Lambda_2,...,\Lambda_l)$. These IHDR include all possible representations of L.

Notice that the basis for $V(\lambda)$ is marked by m non-negative integers, where

$$m = \tfrac{1}{2}(\dim L - \text{rank } L). \tag{4.6}$$

Thus the IHBR obtained from the representation on $V(\lambda)$ uses creation and annihilation operators of m boson states and the corresponding IHDR uses m independent variables. Because $V(\lambda)$ is the irreducible module, m is the minimal number to construct all possible inhomogeneous differential representations. This fact is in accord with Shifman's arguments.

V. IHBR AND IHDR OF SU(3)

We choose the ordered basis for Lie algebra SU(3) as

$$\{e_{31},e_{32},e_{21},\ h_1 = e_{11} - e_{22},\ h_2 = e_{22} - e_{33},\ e_{13},e_{23},e_{12}\}, \tag{5.1}$$

where e_{ij} is a 3×3 matrix with matrix element $(e_{ij})_{kl} = \delta_{ik}\delta_{jl}$. The basis for the universal enveloping algebra $\mathscr{U}(\text{SU}(3))$ of SU(3) can be chosen as

$$\{e_{31}^m e_{32}^n e_{21}^p h_1^k h_2^l e_{13}^r e_{32}^s e_{21}^t | m,n,p,k,l,r,s,t\in\mathbb{Z}^+\}. \tag{5.2}$$

The basis for the standard cyclic module $Z(\lambda) = \mathscr{U}(\text{SU}(3))/I$, where I is a left ideal generated by $\{e_{13},e_{33},e_{12},\ h_1 - \lambda(h_1)1,\ h_2 - \lambda(h_2)1\}$, can be chosen as

$$\{X(m,n,p)\equiv(e_{31}^m e_{32}^n e_{21}^p)\text{Mod } I | m,n,p\in\mathbb{Z}^+\}. \tag{5.3}$$

The representation on $Z(\lambda)$ is obtained as[7]

$$\rho(h_1)X(m,n,p) = (\Lambda_1 + n - m - 2p)X(m,n,p),$$
$$\rho(h_2)X(m,n,p) = (\Lambda_2 - 2n + p - m)X(m,n,p),$$
$$\rho(e_{12})X(m,n,p) = p(\Lambda_1 - p + 1)X(m,n,p-1) - mX(m-1,n+1,p),$$
$$\rho(e_{23})X(m,n,p) = n(\Lambda_2 - m + p - n + 1)X(m,n-1,p) + mX(m-1,n,p+1),$$
$$\rho(e_{13})X(m,n,p) = np(\Lambda_1 - p + 1)X(m,n-1,p-1) + m(\Lambda_1 + \Lambda_2 + p - m - n + 1) \times X(m-1,n,p), \tag{5.4}$$
$$\rho(e_{31})X(m,n,p) = X(m+1,n,p),$$
$$\rho(e_{32})X(m,n,p) = X(m,n+1,p),$$
$$\rho(e_{21})X(m,n,p) = X(m,n,p+1) - nX(m+1,n-1,p),$$
$$[\Lambda_1 = \lambda(h_1), \quad \Lambda_2 = \lambda(h_2)].$$

From the Fock representation that corresponds to (5.4) the IHBR is obtained as

$$\rho(h_1) = \Lambda_1 - a_1^+a_1 + a_2^+a_2 - 2a_3^+a_3,$$
$$\rho(h_2) = \Lambda_2 - a_1^+a_1 - 2a_2^+a_2 + a_3^+a_3,$$
$$\rho(e_{12}) = \Lambda_1a_1 - a_3^+a_3^2 - a_2^+a_1,$$
$$\rho(e_{23}) = \Lambda_2a_2 - a_1^+a_1a_2 + a_3^+a_3a_2 - a_2^+a_2^2,$$

$$\rho(e_{13}) = \Lambda_1 a_2 a_3 - a_3^+ a_3^2 a_2 + (\Lambda_1 + \Lambda_2) a_1 + a_3^+ a_3 a_1 - a_1^+ a_1^2 - a_2^+ a_2 a_1, \quad (5.5)$$

$$\rho(e_{31}) = a_1^+,$$

$$\rho(e_{32}) = a_2^+,$$

$$\rho(e_{31}) = a_3^+ - a_1^+ a_2.$$

The corresponding IHDR is obtained as

$$D(h_1) = \Lambda_1 - \xi\frac{\partial}{\partial\xi} + \eta\frac{\partial}{\partial\eta} - 2\zeta\frac{\partial}{\partial\zeta},$$

$$D(h_2) = \Lambda_2 - \xi\frac{\partial}{\partial\xi} - 2\eta\frac{\partial}{\partial\eta} + \zeta\frac{\partial}{\partial\zeta},$$

$$D(e_{12}) = \Lambda_1\frac{\partial}{\partial\xi} - \zeta\frac{\partial^2}{\partial\zeta^2} - \eta\frac{\partial}{\partial\xi},$$

$$D(e_{23}) = \Lambda_2\frac{\partial}{\partial\eta} - \xi\frac{\partial^2}{\partial\xi\,\partial\eta} + \zeta\frac{\partial^2}{\partial\zeta\,\partial\eta} - \eta\frac{\partial^2}{\partial\eta^2},$$

$$D(e_{13}) = \Lambda_1\frac{\partial^2}{\partial\eta\,\partial\zeta} - \zeta\frac{\partial^2}{\partial\zeta^2}\frac{\partial}{\partial\eta} + (\Lambda_1 + \Lambda_2)\frac{\partial}{\partial\xi} + \zeta\frac{\partial^2}{\partial\zeta\,\partial\xi} - \xi\frac{\partial^2}{\partial\xi^2} - \eta\frac{\partial^2}{\partial\eta\,\partial\xi}, \quad (5.6)$$

$$D(e_{31}) = \xi,$$

$$D(\epsilon_{32}) = \eta,$$

$$D(e_{21}) = \zeta - \xi\frac{\partial}{\partial\eta},$$

which is a realization on the infinite-dimensional space of polynomials

$$\{\xi^m\eta^n\zeta^p | m,n,p\in\mathbb{Z}^+\}. \quad (5.7)$$

If $\Lambda_1,\ \Lambda_2\in\mathbb{Z}^+$, the left ideal $Y(\Lambda_1,\Lambda_2)\equiv Y(\lambda)$ generated by $\{X(0,0,1)^{\Lambda_1+1}, X(0,1,0)^{\Lambda_2+1}\}$ is the maximal proper module and the quotient module $V(\Lambda_1,\Lambda_2) = Z(\lambda)/Y(\Lambda_1,\Lambda_2)$ spanned by

$$\{\tilde{X}(m,n,p)\equiv X(m,n,p)\text{Mod } Y(\Lambda_1,\Lambda_2) | m,n,p\in\mathbb{Z}^+\} \quad (5.8)$$

is the finite-dimensional irreducible module. But it is difficult to determine the linear independent basis for $V(\Lambda_1,\Lambda_2)$ from the set (5.8), and the resulting expressions of the representations on $V(\Lambda_1,\Lambda_2)$ are complicated and unenlightening.[7] However, for some special cases we can give concrete discussions.

A. Triplet

When $\Lambda_1 = 1$, $\Lambda_2 = 0$, the basis for $V(1,0)$ can be chosen as

$$\{\tilde{X}(0,0,0),\tilde{X}(0,0,1),\tilde{X}(0,1,1)\equiv\tilde{X}(1,0,0)\}, \quad (5.9)$$

where $e_{31} = e_{32}e_{21}$ is used. The representation induced on $V(1,0)$ is just the triplet of SU(3). From this triplet we can obtain the IHDR on the space of polynomials with basis.

$$\{1,\zeta,\eta\zeta\}. \quad (5.10)$$

B. Symmetrized IHDR of SU(3)

We start from (5.4).

It is observed that the subspace $J(\Lambda_1,\Lambda_2)$ with basis

$$\{X(m,n,p)\in Z(\lambda) | \Lambda_1\in\mathbb{Z}^+, \Lambda_2\in\mathbb{C}, p\geqslant\Lambda_1+1\} \quad (5.11)$$

is invariant. On the quotient space $W(\Lambda_1,\Lambda_2) \equiv Z(\lambda)/J(\Lambda_1,\Lambda_2)$ with basis

$$\{\overline{X}(m,n,p)\equiv(\Lambda_1 - p)!X(m,n,p)\text{Mod } J(\Lambda_1,\Lambda_2) | m,n\in\mathbb{Z}^+, p\leqslant\Lambda_1\}, \quad (5.12)$$

Eq. (5.3) induces a representation as

$$\rho(h_1)\overline{X}(m,n,p) = (\Lambda_1 + n - m - 2p)\overline{X}(m,n,p),$$

$$\rho(h_2)\overline{X}(m,n,p) = (\Lambda_2 - 2n + p - m)\overline{X}(m,n,p),$$

$$\rho(e_{12})\overline{X}(m,n,p) = p\overline{X}(m,n,p-1) - m\overline{X}(m-1,n+1,p),$$

$$\rho(e_{23})\overline{X}(m,n,p) = n(\Lambda_2 - m + p - n + 1)\overline{X}(m,n-1,p) + m(\Lambda_1 - p)\overline{X}(m-1,n,p+1),$$

$$\rho(e_{13})\overline{X}(m,n,p) = np\overline{X}(m,n-1,p-1) + m(\Lambda_1 + \Lambda_2 + p - n - m + 1)\overline{X}(m-1,n,p), \quad (5.13)$$

$$\rho(e_{31})\overline{X}(m,n,p) = \overline{X}(m+1,n,p),$$

$$\rho(e_{32})\overline{X}(m,n,p) = \overline{X}(m,n+1,p),$$

$$\rho(e_{21})\overline{X}(m,n,p) = (\Lambda_1 - p)\overline{X}(m,n,p+1) - n\overline{X}(m+1,n-1,p).$$

The IHDR is obtained as

$$D(h_1) = \Lambda_1 - \xi\frac{\partial}{\partial\xi} + \eta\frac{\partial}{\partial\eta} - 2\zeta\frac{\partial}{\partial\zeta}$$

$$D(h_2) = \Lambda_2 - \xi\frac{\partial}{\partial\xi} - 2\eta\frac{\partial}{\partial\eta} + \zeta\frac{\partial}{\partial\zeta}$$

$$D(e_{12}) = \frac{\partial}{\partial\zeta} - \eta\frac{\partial}{\partial\xi}$$

$$D(e_{23}) = \Lambda_2\frac{\partial}{\partial\eta} - \xi\frac{\partial^2}{\partial\xi\,\partial\eta} - \zeta\frac{\partial^2}{\partial\zeta\,\partial\eta} - \eta\frac{\partial^2}{\partial\eta^2} + \Lambda_1\frac{\partial}{\partial\xi} - \zeta\frac{\partial}{\partial\xi} \quad (5.14)$$

$$D(e_{13}) = \frac{\partial^2}{\partial\eta\partial\zeta} + (\Lambda_1 + \Lambda_2)\frac{\partial}{\partial\xi} + \zeta\frac{\partial^2}{\partial\zeta\,\partial\xi} - \eta\frac{\partial^2}{\partial\eta\,\partial\xi} - \xi\frac{\partial^2}{\partial\xi^2}$$

$$D(e_{31}) = \xi, \quad D(e_{32}) = \eta, \quad D(e_{21}) = \Lambda_1 - \zeta - \xi\frac{\partial}{\partial\eta},$$

which is a realization on the space of polynomials

$$\{\xi^m\eta^n\zeta^p | m,n,p\in\mathbb{Z}^+, 0\leqslant p\leqslant\Lambda_1\}. \quad (5.15)$$

Particularly, when $\Lambda_1 = 0$, and $p = 0$, the representation (5.13) becomes

$$\rho(h_1)\overline{X}(m,n) = (n - m)\overline{X}(m,n),$$

$$\rho(h_2)\overline{X}(m,n) = (\Lambda_2 - m - 2n)\overline{X}(m,n),$$

$$\rho(e_{23})\overline{X}(m,n) = n(\Lambda_2 - m - n + 1)\overline{X}(m,n-1),$$

$$\rho(e_{13})\overline{X}(m,n) = m(\Lambda_2 - m - n + 1)\overline{X}(m-1,n),$$

$$\rho(e_{31})\overline{X}(m,n) = \overline{X}(m+1,n),$$

$$\rho(e_{32})\overline{X}(m,n) = \overline{X}(m,n+1),$$

$$\rho(e_{21})\overline{X}(m,n) = -n\overline{X}(m+1,n-1), \quad (5.16)$$

where $\overline{X}(m,n)\equiv\overline{X}(m,n,0)$.

When $\Lambda_2 \in \mathbb{Z}^+$, it is easy to see that the subspace $Q(0,\Lambda_2)$ spanned by

$$\{\overline{X}(m,n) \in W(0,\Lambda_2) | m+n \geqslant \Lambda_2 + 1, m,n \in \mathbb{Z}^+\} \quad (5.17)$$

is invariant. On the quotient space $H(\Lambda_2) = W(0,\Lambda_2)/Q(0,\Lambda_2)$ with basis

$$\{H(m,n) \equiv (\Lambda_2 - m - n)!\overline{X}(m,n) \text{Mod } Q(0,\Lambda_2) | 0 \leqslant m+n \leqslant \Lambda_2\} \quad (5.18)$$

$$\dim H(\Lambda_2) = \tfrac{1}{2}(\Lambda_2+1)(\Lambda_2+2),$$

(5.16) induces a representation on $H(\Lambda_2)$ as

$$\begin{aligned}
&\rho(h_1)H(m,n) = (n-m)H(m,n),\\
&\rho(h_2)H(m,n) = (\Lambda_2 - m - 2n)H(m,n),\\
&\rho(e_{12})H(m,n) = -mH(m-1,n+1),\\
&\rho(e_{23})H(m,n) = nH(m,n-1),\\
&\rho(e_{13})H(m,n) = mH(m-1,n),\\
&\rho(e_{31})H(m,n) = (\Lambda_2 - m - n)H(m+1,n),\\
&\rho(e_{32})H(m,n) = (\Lambda_2 - m - n)H(m,n+1),\\
&\rho(e_{21})H(m,n) = -nH(m+1,n-1).
\end{aligned} \quad (5.19)$$

From the Fock representation that corresponds to (5.19) the IHBR is obtained as

$$\begin{aligned}
&\rho(h_1) = a_2^+ a_2 - a_1^+ a_1, \quad \rho(h_2) = \Lambda_2 - a_1^+ a_1 - 2a_2^+ a_2,\\
&\rho(e_{12}) = -a_2^+ a_1, \quad \rho(e_{23}) = a_2,\\
&\rho(e_{13}) = a_1, \quad \rho(e_{31}) = \Lambda_2 a_1^+ - a_1^{+2} a_1 - a_1^+ a_2^+ a_2,\\
&\rho(e_{32}) = \Lambda_2 a_2^+ - a_2^+ a_1^+ a_1 - a_2^{+2} a_2, \quad \rho(e_{21}) = -a_1^+ a_2.
\end{aligned} \quad (5.20)$$

The corresponding IHDR is obtained as

$$\begin{aligned}
&D(h_1) = \eta\frac{\partial}{\partial\eta} - \xi\frac{\partial}{\partial\xi}, \quad \rho(h_2) = \Lambda_2 - \xi\frac{\partial}{\partial\xi} - 2\eta\frac{\partial}{\partial\eta},\\
&D(e_{12}) = -\eta\frac{\partial}{\partial\xi}, \quad D(e_{23}) = \frac{\partial}{\partial\eta},\\
&D(e_{13}) = \frac{\partial}{\partial\xi}, \quad D(e_{31}) = \Lambda_2\xi - \xi^2\frac{\partial}{\partial\xi} - \xi\eta\frac{\partial}{\partial\eta},\\
&D(e_{32}) = \Lambda_2\eta - \eta\xi\frac{\partial}{\partial\xi} - \eta^2\frac{\partial}{\partial\eta}, \quad D(e_{21}) = -\xi\frac{\partial}{\partial\eta},
\end{aligned} \quad (5.21)$$

which is a realization on the finite dimensional space of polynomials with basis

$$\{\xi^m\eta^n | 0 \leqslant m+n \leqslant \Lambda_2, m,n \in \mathbb{Z}^+\}. \quad (5.22)$$

In fact, if we choose the representation D in Ref. 5 as the antitriplet, we just obtain the realization (5.21) by making use of the general formula in Ref. 5. Therefore, the realization (5.21) marked by a non-negative integer is commensurate with the symmetrized direct product of Λ_2 antitriplets of SU(3). So we call the realization (5.21) the symmetrized IHDR.

C. Antitriplet

Let $\Lambda_2 = 1$. Then we obtain the antitriplet from (5.21) on the three dimensional space of polynomials with basis

$$\{\xi, \eta, 1\}. \quad (5.23)$$

ACKNOWLEDGMENT

The authors would like to thank Professor Zhaoyan Wu, Jilin University, for his useful discussions.

[1] A. V. Turbiner and A. G. Ushiveridze, Phys. Lett. A **126**, 181 (1987).
[2] A. V. Turbiner, Commun. Math. Phys. **118**, 467 (1988).
[3] M. A. Shifman and A. V. Turbiner, Preprint ITEP-174, 1988 and Commun. Math. Phys. (to be published).
[4] M. A. Shifman, Inter. J. Mod. Phys. A. **4**, 2897 (1989).
[5] H. C. Fu and C. P. Sun, J. Math. Phys. **31** 287 (1990).
[6] B. Gruber and A. U. Klimyk, J. Math. Phys. **25**, 755 (1984).
[7] B. Gruber, A. U. Klimyk, and Y. F. Smirnov, Nuovo Cimento A **69**, 97 (1982).
[8] B. Gruber and R. Lenczewski, J. Phys. A **16**, 3703 (1983).
[9] R. Lenczewski and B. Gruber, J. Phys. A **19**, 1 (1986).
[10] J. E. Humphreys, *Introduction to Lie Algebras and Representation Theory* (Springer-Verlag, Berlin, 1972), 2nd printing revised.
[11] H. D. Doebner and B. Gruber, J. Math. Phys. **30**, 594 (1989).

Branching rules for a class of typical and atypical representations of gl($m|n$)

M. D. Gould
Department of Mathematics, The University of Queensland, Brisbane, 4072, Australia

P. D. Jarvis
Department of Physics, The University of Tasmania, Hobart, 7000, Australia

A. J. Bracken
Department of Mathematics, The University of Queensland, Brisbane, 4072, Australia

(Received 1 March 1990; accepted for publication 5 July 1990)

The irreducible representations of the Lie superalgebra gl($m|n$) with highest weights of the form $\Lambda = (\lambda_1,\lambda_2,...\lambda_m|\dot{\omega})$ are investigated using a recently introduced induced module construction for atypical modules. The gl($m|n$)$\downarrow$gl($m|n-1$) branching rules are obtained and a suitable Gel'fand–Tsetlin basis is introduced. The class of representations considered includes some multiply atypical irreducible representations of gl($m|n$) and all irreducible representations of gl($m|1$).

I. INTRODUCTION

The concept of supersymmetry has applications in a variety of areas,[1,2] including condensed matter physics[3–5] and nuclear physics,[6] as well as particle physics, where it was first introduced.[7] Underlying this concept is the theory of Lie superalgebras which, in the case of the so-called "basic classical" superalgebras and their finite-dimensional irreducible representations, has largely been developed by Kac.[8,9] Of all the Lie superalgebras, it is the basic classical superalgebras whose theory is most like that of simple Lie algebras.

However, there are crucial differences that have resulted in many aspects of their representation theory remaining only partly explored. Kac[9] showed that there are two disjoint classes of finite-dimensional irreducible representations of any basic classical superalgebra, and named these classes typical and atypical. Typical representations have many properties in common with finite-dimensional representations of simple Lie algebras, and in particular can be given explicitly by an induced module construction that allows a straightforward determination of their characters and dimensions.[9] On the other hand, the situation with atypical representations is far more complex and they are still not well understood, although progress has been made in various directions.[10–31]

In this paper we are concerned with finite-dimensional irreducible representations of the Lie superalgebra gl($m|n$), which we work with in preference to its basic classical subalgebra sl($m|n$). The matrix elements of the generators in an orthonormal Gel'fand–Tsetlin basis have already been obtained, for each finite-dimensional irreducible representation of sl($m|1$) and gl($m|1$), by Palev.[25,26] More recently the branching rules and characters of all finite-dimensional irreducible representations of gl($m|1$) have been determined by Gould, Bracken, and Hughes[28] using the modified induced module construction recently introduced for atypical representations by Gould.[27] The latter results confirm, for gl($m|1$), the character formula conjectured by Hughes and King,[22] as distinct from the formula obtained previously by Bernstein and Leites[10] and by Van der Jeugt.[23] More recently, all finite-dimensional irreducible star and grade star representations of gl($m|1$) have been classified.[29]

Despite this recent progress in the representation theory of gl($m|1$), the finite-dimensional irreducible atypical representations of gl($m|n$) for $n>1$ are still poorly understood. One of the complicating factors is that gl($m|n$) for $m,n>1$ admits multiply atypical irreducible representations, whereas gl($m|1$) only has singly atypical irreducible representations (in the terminology used by Hughes and King[22]). Another factor is that an irreducible representation of gl($m|n$), when regarded as a representation of the gl($m|k$) subalgebra ($0<k<n$), is not in general completely reducible. In this connection Palev[30] has considered a class of "essentially typical" representations of gl($m|n$) that, by definition, do not present this second difficulty.

It is our aim in this paper to investigate the structure of a different class of finite-dimensional irreducible representations of gl($m|n$), those with highest weights of the special form $\Lambda = (\lambda_1,\lambda_2,...,\lambda_m|\omega,\omega,...,\omega) = (\lambda|\dot{\omega})$, where λ is the highest weight of an arbitrary finite-dimensional irreducible representation of the Lie algebra gl(m), and ω is an arbitrary complex number. These representations are of special interest for physics because amongst them are the representations that arise in the interacting boson–fermion model of the nucleus.[6,32] We note that this class in general contains multiply atypical representations of gl($m|n$) (indeed, for $m\leqslant n$, representations with maximum atypicality occur) and in the case of gl($m|1$), contains all finite-dimensional irreducible representations, so it certainly contains representations that are not essentially typical.

Nevertheless, we shall show with the help of the modified induced module construction,[27] the surprising result that any representation of gl($m|n$) of the type described is for $n>1$, a direct sum of irreducible representations of gl($m|n-1$) of the same type. The corresponding gl($m|n$)$\downarrow$gl($m|n-1$) [and gl($m|n$)$\downarrow$gl(m)$\oplus$gl(n)] branching rules will be obtained and a suitable orthogonal Gel'fand–Tsetlin basis introduced. It would clearly be of interest in future work to determine the matrix elements of the generators in such a basis (cf. Palev[30]).

 0022-2488/90/122803-08$03.00

II. PRELIMINARIES

The generators of the Lie superalgebra $\mathrm{gl}(m|n)$ are given by the even $\mathrm{gl}(m)\oplus\mathrm{gl}(n)$ generators $E^i_j\,(1\leqslant i,j\leqslant m)$ and $E^\mu_\nu\,(1\leqslant\mu,\nu\leqslant n)$, respectively, satisfying the usual commutation relations, together with the odd generators ψ^i_μ, and $\psi^\mu_i\,(1\leqslant i\leqslant m,\ 1\leqslant\mu\leqslant n)$ satisfying the commutation and anticommutation relations:

$$[E^i_j,\psi^k_\mu]_- = \delta^k_j\psi^i_\mu,\quad [E^\sigma_\nu,\psi^k_\mu]_- = -\delta^\sigma_\mu\psi^k_\nu, \tag{1a}$$

$$[E^i_j,\psi^\mu_k]_- = -\delta^i_k\psi^\mu_j,\quad [E^\sigma_\nu,\psi^\mu_k]_- = \delta^\mu_\nu\psi^\sigma_k, \tag{1b}$$

$$[\psi^i_\mu,\psi^j_\nu]_+ = [\psi^\mu_i,\psi^\nu_j]_+ = 0,\quad [\psi^i_\mu,\psi^\nu_j]_+ = \delta^\nu_\mu E^i_j + \delta^i_j E^\nu_\mu, \tag{1c}$$

where $[,]_-$ (resp. $[,]_+$) denotes the commutator (resp. anticommutator): these two cases are taken into account below by the graded bracket, denoted $[,]$. We note that (1a) [resp (1b)] expresses the fact that the generators ψ^i_μ (resp. ψ^μ_i) transform as the irreducible representation $(1,\dot 0)\otimes(\dot 0,-1)$ [resp. $(\dot 0,-1)\otimes(1,\dot 0)$] of the Lie algebra $\mathrm{gl}(m)\oplus\mathrm{gl}(n)$.

A basis for a Cartan subalgebra of $\mathrm{gl}(m|n)$ consists of the commuting operators $E^i_i\,(1\leqslant i\leqslant m), E^\mu_\mu\,(1\leqslant\mu\leqslant n)$ whose eigenvalues serve to label the weights of the representations. We denote the weights Λ of $\mathrm{gl}(m|n)$ by (notation as in Kac[8])

$$\Lambda = \sum_{i=1}^m \Lambda^{(0)}_i\epsilon_i + \sum_{\mu=1}^n \Lambda^{(1)}_\mu\delta_\mu = (\Lambda^{(0)}_1,\ldots,\Lambda^{(0)}_m|\Lambda^{(1)}_1,\ldots,\Lambda^{(1)}_n), \tag{2}$$

so that, with this convention, the root system of $\mathrm{gl}(m|n)$ is given by the set of even roots

$$\pm(\epsilon_i-\epsilon_j),\quad 1\leqslant i<j\leqslant m,\quad \pm(\delta_\mu-\delta_\nu),\quad 1\leqslant\mu<\nu\leqslant n$$

together with the set of odd roots

$$\pm(\epsilon_i-\delta_\mu),\quad 1\leqslant i\leqslant m,\quad 1\leqslant\mu\leqslant n.$$

Following Kac[8] we choose, as a system of simple roots, the distinguished set

$$\epsilon_i-\epsilon_{i+1}\,(1\leqslant i<m),\quad \alpha_s=\epsilon_m-\delta_1,$$
$$\delta_\mu-\delta_{\mu+1}\,(1\leqslant\mu<n)$$

so that the sets of even and odd positive roots are given, respectively, by

$$\Phi_0^+ = \{\epsilon_i-\epsilon_j|1\leqslant i<j\leqslant m\}\cup\{\delta_\mu-\delta_\nu|1\leqslant\mu<\nu\leqslant n\},$$
$$\Phi_1^+ = \{\epsilon_i-\delta_\mu|1\leqslant i\leqslant m,1\leqslant\mu\leqslant n\}$$

and we set

$$\rho_0 = \frac{1}{2}\sum_{\alpha\in\Phi_0^+}\alpha = \frac{1}{2}\sum_{i=1}^m(m+1-2i)\epsilon_i + \frac{1}{2}\sum_{\mu=1}^n(n+1-2\mu)\delta_\mu,$$

$$\rho_1 = \frac{1}{2}\sum_{\alpha\in\Phi_1^+}\alpha = \frac{n}{2}\sum_{i=1}^m\epsilon_i - \frac{m}{2}\sum_{\mu=1}^n\delta_\mu = \frac{1}{2}(\dot n|-\dot m),$$

$$\rho=\rho_0-\rho_1.$$

We note that $\mathrm{gl}(m|n)$ admits a nondegenerate even invariant bilinear supertrace form arising from the fundamental vector representation π:

$$(x,y) = \mathrm{str}(\pi(x)\pi(y)),\quad x,y\in\mathrm{gl}(m|n)$$

leading to

$$(\epsilon_i,\epsilon_j)=\delta_{ij},\quad (\epsilon_i,\delta_\mu)=0,\quad (\delta_\mu,\delta_\nu)=-\delta_{\mu\nu}.$$

This in turn induces a nondegenerate bilinear form on the weights, given by

$$(\Lambda,\Lambda') = \sum_{i=1}^m\Lambda^{(0)}_i\Lambda'^{(0)}_i - \sum_{\mu=1}^n\Lambda^{(1)}_\mu\Lambda'^{(1)}_\mu \tag{3}$$

with Λ as in Eq. (2) and $\Lambda' = (\Lambda'^{(0)}|\Lambda'^{(1)})$.

Every finite-dimensional irreducible $\mathrm{gl}(m|n)$ module admits a unique (up to scalar multiples) $\mathrm{gl}(m)\oplus\mathrm{gl}(n)$ highest weight vector v^Λ of weight Λ satisfying the conditions

$$\psi^i_\mu v^\Lambda = 0,\quad 1\leqslant i\leqslant m,\quad 1\leqslant\mu\leqslant n.$$

Such a vector is called a highest weight vector and its weight Λ, called the highest weight of the representation, uniquely characterizes the representation (Kac[8]): throughout we denote the irreducible $\mathrm{gl}(m|n)$ module with highest weight $\Lambda=(\Lambda^{(0)}|\Lambda^{(1)})$ by $V(\Lambda)$. The components of the highest weight Λ necessarily satisfy the conditions

$$\Lambda^{(0)}_i-\Lambda^{(0)}_j\in\mathbb{Z}^+\,(1\leqslant i<j\leqslant m),$$
$$\Lambda^{(1)}_\mu-\Lambda^{(1)}_\nu\in\mathbb{Z}^+\,(1\leqslant\mu<\nu\leqslant n), \tag{4}$$

which are just the conditions that Λ constitute a dominant integral weight of the Lie algebra $\mathrm{gl}(m)\oplus\mathrm{gl}(n)$: throughout we denote the finite-dimensional irreducible module over $\mathrm{gl}(m)\oplus\mathrm{gl}(n)$ with highest weight Λ by $V_0(\Lambda)$. We denote the set of all weights Λ, whose component satisfy (4), by D^+ (the set of dominant integral weights).

Corresponding to every $\Lambda\in D^+$ we may construct an indecomposable finite dimensional $\mathrm{gl}(m|n)$ module with highest weight Λ using the induced module construction of Kac[8]. To this end we find it convenient for the moment to denote the Lie superalgebra $\mathrm{gl}(m|n)$ simply by L and to let L_0 denote the Lie subalgebra $\mathrm{gl}(m)\oplus\mathrm{gl}(n)$. We let L_+ (resp. L_-) denote the graded Alebian subalgebra spanned by the operators ψ^i_μ (resp. ψ^μ_i) giving the $\mathbb{Z}_2$-consistent $\mathbb{Z}$-gradation

$$L = L_-\oplus L_0\oplus L_+. \tag{5}$$

We denote the universal enveloping algebras of L, L_0, $L_\pm$, by U, U_0, $U_\pm$ resp. and we denote by $\overline{U}_\pm$ the universal enveloping algebra of the subalgebra

$$\overline{L}_\pm = L_0\oplus L_\pm. \tag{6}$$

We note that the algebras $U_\pm$ are 2^{mn}-dimensional with basis consisting of $1\in\mathbb{C}$ together with all basis monomials

$$\psi^{i_1}_{\mu_1}\psi^{i_2}_{\mu_2}\cdots\psi^{i_k}_{\mu_k}\ (\text{resp.},\psi^{\mu_1}_{i_1}\psi^{\mu_2}_{i_2}\cdots\psi^{\mu_k}_{i_k}),$$
$$1\leqslant\mu_1\leqslant\mu_2\leqslant\cdots\leqslant\mu_k\leqslant n, \tag{7}$$

where $1\leqslant i_1,i_2,\cdots,i_k\leqslant m$ subject to the condition that, for $\mu_r=\mu_l\,(r<l)$, we have $1\leqslant i_r<\cdots<i_l\leqslant m$. From the Poincare–Birkhoff–Witt theorem[31] we may write, in view of (5), (6):

$$U = U_-U_0U_+ = U_+U_0U_- = U_-\overline{U}_+ = U_+\overline{U}_-. \tag{8}$$

Given a finite-dimensional irreducible L_0 module $V_0(\Lambda)$, we turn $V_0(\Lambda)$ into a $\overline{U}_+$ module by defining

$$L_+ V_0(\Lambda) = (0). \tag{9}$$

The induced L-module $\overline{V}(\Lambda)$ is then defined by[8]

$$\overline{V}(\Lambda) = U \otimes_{\overline{U}_+} V_0(\Lambda) \tag{10}$$

which is spanned by all vectors of the form [notation as in Eq. (7)]

$$\psi^{\mu_1}_{i_1}\psi^{\mu_2}_{i_2}\cdots\psi^{\mu_k}_{i_k} \otimes v, \quad v \in V_0(\Lambda).$$

We note that the induced module (10) is an indecomposable L module with highest weight Λ and dimension $\dim \overline{V}(\Lambda) = 2^{mn} \dim V_0(\Lambda)$. In a similar way we may define

$$L_- V_0(\Lambda) = (0), \tag{11}$$

which leads to the induced L module

$$\overline{V}_-(\Lambda) = U \otimes_{\overline{U}_-} V_0(\Lambda). \tag{12}$$

This is also indecomposable, but in this case is cyclically generated by a lowest weight vector of weight Λ_-, where Λ_- is the lowest weight of $V_0(\Lambda)$: recall that

$$\Lambda_- = (\Lambda^{(0)}_m, \Lambda^{(0)}_{m-1}, \cdots, \Lambda^{(0)}_1 | \Lambda^{(1)}_n, \Lambda^{(1)}_{n-1}, \cdots, \Lambda^{(1)}_1).$$

We have the following result.

Theorem 1: (Kac[8])

$\overline{V}(\Lambda) = V(\Lambda)$ (i.e., $\overline{V}(\Lambda)$ is irreducible) if and only if $(\Lambda + \rho, \alpha) \neq 0$, for all $\alpha \in \Phi_1^+$. ■

A weight $\Lambda \in D^+$ satisfying the conditions of Theorem 1, and the corresponding module $V(\Lambda)$, are said to be *typical*. Similarly, Λ and $V(\Lambda)$ are called *atypical* if $(\Lambda + \rho, \alpha) = 0$ for some $\alpha \in \Phi_1^+$; we note that

$$(\Lambda + \rho, \epsilon_i - \delta_\mu) = \Lambda^{(0)}_i + \Lambda^{(1)}_\mu + m + 1 - i - \mu. \tag{13}$$

Thus in the case that $V(\Lambda)$ is a typical gl$(m|n)$ module, and so equal to $\overline{V}(\Lambda)$, the structure of $V(\Lambda)$ follows immediately from the induced module construction (10). In the case that $V(\Lambda)$ is atypical, however, the situation is more complex, since in such a case the Kac module $\overline{V}(\Lambda)$ is no longer irreducible and $V(\Lambda)$ only appears as a subquotient of $\overline{V}(\Lambda)$:

$$V(\Lambda) \cong \overline{V}(\Lambda)/M(\Lambda)$$

where $M(\Lambda)$ is the (unique) maximal submodule of $\overline{V}(\Lambda)$. However for arbitrary $\Lambda \in D^+$, we may employ the result[27] that the lowest weight Kac module [c.f. Eqs. (11), (12)]

$$\overline{V}_-(\Lambda - 2\rho_1)$$
$$= U \otimes_{\overline{U}_-} V_0(\Lambda - 2\rho_1),\ L_- V_0(\Lambda - 2\rho_1) = (0)$$

contains, as a unique irreducible L module,

$$V(\Lambda) = U[\bar{\psi} \otimes V_0(\Lambda - 2\rho_1)]$$
$$= U_-[\bar{\psi} \otimes V_0(\Lambda - 2\rho_1)], \tag{14}$$

where

$$\bar{\psi} = \prod_{\mu=1}^{n} \prod_{i=1}^{m} \psi^i_\mu. \tag{15}$$

Since the ψ^i_μ anticommute, the operator (15) is uniquely determined up to a sign: this operator is denoted T_+ in the work of Kac.[8] We note, since $(\rho_1, \alpha) = 0$ for $\alpha \in \Phi_0^+$, that $\Lambda \in D^+$ if and only if $\Lambda - 2\rho_1 \in D^+$.

It is our aim in the following to employ the modified induced module construction of Eq. (14) to investigate the irreducible gl$(m|n)$ modules. In particular we determine the gl$(m|n) \downarrow$gl$(m|n-1)$ branching rules ($n > 1$) for the irreducible representations with highest weights $\Lambda = (\Lambda^{(0)}|\Lambda^{(1)})$ whose odd component is trivial: $\Lambda^{(1)} = (\dot{\omega})$.

III. INDUCED MODULE CONSTRUCTION

Following Ref. 27 (see also Ref. 28) we introduce the tensors $\bar{\psi}^\mu_i$ defined by

$$\psi^j_\nu \bar{\psi}^\mu_i = \delta^\mu_\nu \delta^j_i \bar{\psi},$$

which transform as the irreducible representation (irrep.) of gl$(m) \oplus$ gl(n) with highest weight $(\dot{0}, -1|1,\dot{0}) - 2\rho_1$:

$$[E^i_j, \bar{\psi}^\mu_k] = n\delta^i_j \bar{\psi}^\mu_k - \delta^i_k \bar{\psi}^\mu_j,$$
$$[E^\mu_\nu, \bar{\psi}^\sigma_k] = \delta^\sigma_\nu \bar{\psi}^\mu_k - m\delta^\mu_\nu \bar{\psi}^\sigma_k. \tag{16}$$

We similarly introduce the antisymmetric tensors

$$\bar{\psi}^{\mu\nu}_{ij} = -\bar{\psi}^{\nu\mu}_{ij}$$

defined by[27]

$$\psi^i_\mu \bar{\psi}^{\nu\tau}_{jk} = \delta^i_j \delta^\nu_\mu \bar{\psi}^\tau_k - \delta^\tau_\mu \delta^i_k \bar{\psi}^\nu_j.$$

More generally we have the antisymmetric tensors

$$\bar{\psi}^{\mu_1\cdots\mu_k}_{i_1\cdots i_k} = sn\pi \bar{\psi}^{\mu_{\pi(1)}\cdots\mu_{\pi(k)}}_{i_1\cdots i_k}$$

(π any permutation of numbers $1,\ldots,k$) which are defined recursively by

$$\psi^i_\mu \bar{\psi}^{\nu_1\cdots\nu_k}_{j_1\cdots j_k} = \delta^{\nu_1}_\mu \delta^i_{j_1} \bar{\psi}^{\nu_2\cdots\nu_k}_{j_2\cdots j_k} - \delta^{\nu_2}_\mu \delta^i_{j_2} \bar{\psi}^{\nu_1\nu_3\cdots\nu_k}_{j_1 j_3\cdots j_k}$$
$$+ \cdots + (-1)^{k-1} \delta^{\nu_k}_\mu \delta^i_{j_k} \psi^{\nu_1\cdots\nu_{k-1}}_{j_1\cdots j_{k-1}}. \tag{17}$$

It follows from their definition that the above tensors satisfy the commutation relations [cf. Eq. (16)]

$$\left[E^i_j, \bar{\psi}^{\mu_1\cdots\mu_k}_{j_1\cdots j_k}\right] = n\delta^i_j \bar{\psi}^{\mu_1\cdots\mu_k}_{j_1\cdots j_k} - \delta^i_{j_1} \bar{\psi}^{\mu_1\mu_2\cdots\mu_k}_{ij_2\cdots j_k} - \cdots$$
$$- \delta^i_{j_k} \bar{\psi}^{\mu_1\cdots\mu_k}_{j_1\cdots j} \tag{18a}$$

$$\left[E^\mu_\nu, \bar{\psi}^{\mu_1\cdots\mu_k}_{j_1\cdots j_k}\right] = = \delta^{\mu_1}_\nu \bar{\psi}^{\mu\mu_2\cdots\mu_k}_{j_1 j_2\cdots j_k} + \cdots + \delta^{\mu_k}_\nu \bar{\psi}^{\mu_1\cdots\mu_{k-1}\mu}_{j_1\cdots j_{k-1} j_k}$$
$$- m\delta^\mu_\nu \bar{\psi}^{\mu_1\cdots\mu_k}_{j_1\cdots j_k}. \tag{18b}$$

As well we note the following graded commutation relations that may be proved using the methods of Refs. 27 and 28:

$$\left[\psi^\mu_i, \bar{\psi}^{\nu_1\cdots\nu_k}_{j_1\cdots j_k}\right] = E^q_i \bar{\psi}^{\mu\nu_1\cdots\nu_k}_{qj_1\cdots j_k} + \bar{\psi}^{\sigma\nu_1\cdots\nu_k}_{ij_1\cdots j_k} E^\mu_\sigma \tag{19}$$

(summation over $1 \leqslant q \leqslant m$, $1 \leqslant \sigma \leqslant n$, implied). In particular we obtain

$$[\psi^\mu_i, \bar{\psi}] = E^q_i \bar{\psi}^\mu_q + \bar{\psi}^\sigma_i E^\mu_\sigma, \tag{20a}$$

$$[\psi^\mu_i, \bar{\psi}^\nu_j] = E^q_i \bar{\psi}^{\mu\nu}_{qj} + \bar{\psi}^{\sigma\nu}_{ij} E^\mu_\sigma. \tag{20b}$$

To construct the irreducible gl$(m|n)$ module with highest weight $\Lambda = (\lambda|\dot{\omega})$ we introduce the irreducible gl$(m) \oplus$ gl(n) module $V_0(\Lambda - 2\rho_1)$ satisfying

$$L_- V_0(\Lambda - 2\rho_1) = (0).$$

Equation (14) then implies that the irreducible gl$(m|n)$ module $V(\Lambda)$ is given by

$$V(\Lambda) = U_- \bar{\psi} \otimes V_0(\Lambda - 2\rho_1),$$

which is spanned by vectors

$$\psi^{\mu_1}_{i_1}\cdots\psi^{\mu_k}_{i_k}\bar{\psi} \otimes w, \quad w \in V_0(\Lambda - 2\rho_1).$$

We note that for the special irreps we are considering the representation labels of $\Lambda' = \Lambda - 2\rho_1$ are given by

$$\Lambda_i^{(0)'} = \lambda_i - n, \quad 1 \leqslant i \leqslant m, \quad \Lambda_\mu^{(1)'} = \omega + m, \quad 1 \leqslant \mu \leqslant n,$$

so that, for $w \in V_0(\Lambda - 2\rho_1)$,

$$E_\nu^\mu w = (\omega + m)\delta_\nu^\mu w. \tag{21}$$

Now we have, from Eqs. (20a, 21), for all $w \in V_0(\Lambda - 2\rho_1)$

$$\psi_i^\mu \bar{\psi} \otimes w = [\psi_i^\mu, \bar{\psi}] \otimes w$$
$$= (\omega + m - \overline{E})_i^q \bar{\psi}_q^\mu \otimes w \tag{22a}$$

where we have adopted the convention

$$(\omega + m - \overline{E})_i^j = (\omega + m)\delta_i^j - \overline{E}_i^j$$

and $\overline{E}$ denotes the gl(m) adjoint matrix defined by[34]

$$\overline{E}_i^j = -E_i^j.$$

More generally using Eq. (19) we obtain

$$\psi_i^\mu \bar{\psi}_{j_1 \cdots j_k}^{\nu_1 \cdots \nu_k} \otimes w = (\omega + m - \overline{E})_i^q \bar{\psi}_{qj_1 \cdots j_k}^{\mu\nu_1 \cdots \nu_k} \otimes w. \tag{22b}$$

Following Bracken and Green[33,34] the gl(m) $\oplus$ gl(n) tensor operator ψ_i^μ may be resolved into gl(m) shift components

$$\psi_i^\mu = \sum_{r=1}^{m} \psi[r]_i^\mu,$$

which effect the following shifts on the gl(m) $\oplus$ gl(n) representation labels:

$$\psi[r]_i^\mu v \in V_0(\Lambda - \epsilon_r + \delta_1), \quad v \in V_0(\Lambda), \quad \Lambda = (\lambda \,|\dot{\omega}).$$

The above shift components may be constructed using the projection operators[34]

$$\overline{P}[r] = \prod_{l \neq r}^{m} \left(\frac{\overline{E} - \overline{\alpha}_l}{\overline{\alpha}_r - \overline{\alpha}_l} \right)$$

according to[34]

$$\psi[r]_i^\mu = \overline{P}[r]_i^j \psi_j^\mu,$$

where the adjoint roots $\overline{\alpha}_r$ take constant values on an irreducible gl(m) module with highest weight λ given by

$$\overline{\alpha}_r = r - 1 - \lambda_r. \tag{23}$$

Alternatively in terms of the gl(m) vector matrix[34] $E = (E_j^i)$ we may write

$$\psi[r]_i^\mu = \psi_j^\mu P[r]_i^j,$$

where $P[r]$ denotes the projection[33,34]

$$P[R] = \prod_{l \neq r}^{m} \left(\frac{E - \alpha_l}{\alpha_r - \alpha_l} \right)$$

and the roots α_r are related to the adjoint roots $\overline{\alpha}_r$ by $\alpha_r = n - 1 - \overline{\alpha}_r$.

Using Eq. (22a) we have, for the action of the above shift components,

$$\psi[r]_i^\mu \bar{\psi} \otimes w = \overline{P}[r]_i^j \psi_j^\mu \bar{\psi} \otimes w$$
$$= \overline{P}[r]_i^j (\omega + m - \overline{E})_j^q \bar{\psi}_q^\mu \otimes w$$
$$= (\omega + m - \overline{\alpha}_r) \bar{\psi}[r]_i^\mu \otimes w, \tag{24}$$

where

$$\bar{\psi}[r]_i^\mu = \overline{P}[r]_i^j \bar{\psi}_j^\mu = \bar{\psi}_j^\mu P[r]_i^j$$

and we have employed the gl(m) characteristic identity[34]

$$\overline{E}_i^j \overline{P}[r]_j^k = \overline{\alpha}_r \overline{P}[r]_i^k. \tag{25}$$

We note that the adjoint roots $\overline{\alpha}_r$ are given by [cf. Eq. (23)] $\overline{\alpha}_r = r - 1 - \lambda'_r$ where λ' now refers to the shifted weight $\lambda' = \lambda - \epsilon_r$. Using

$$(\Lambda + \rho, \epsilon_r - \delta_\mu)$$
$$= \lambda_r + \omega + m + 1 - r - \mu, \quad \text{for } \Lambda = (\lambda \,|\dot{\omega}) \tag{26}$$

we have

$$(\Lambda + \rho, \epsilon_r - \delta_1) = \omega + m - r + \lambda_r$$
$$= \omega + m - \overline{\alpha}_r$$

so that Eq. (24) may be expressed as

$$\psi[r]_i^\mu \bar{\psi} \otimes w = (\Lambda + \rho, \epsilon_r - \delta_1) \bar{\psi}[r]_i^\mu \otimes w. \tag{27}$$

The states (27) either span the irreducible gl(m) $\oplus$ gl(n) submodule of $V(\Lambda)$ with highest weight $\Lambda + \epsilon_r + \delta_1$ or else are zero for all $w \in V_0(\Lambda - 2\rho_1)$ and choice of indices μ, i. It follows that the irreducible gl(m) $\oplus$ gl(n) module $V_0(\Lambda - \epsilon_r + \delta_1)$ occurs (with unit multiplicity) in $V(\Lambda)$ if and only if $\Lambda - \epsilon_r + \delta_1$ is dominant and $(\Lambda + \rho, \epsilon_r - \delta_1) \neq 0$.

Proceeding further we now consider the action of two odd lowering operators $\psi_i^\mu \psi_j^\nu$ on the vectors $\bar{\psi} \otimes w$. We have, using Eqs. (22b) and (27)

$$\psi_i^\mu \psi[r]_j^\nu \bar{\psi} \otimes w = (\Lambda + \rho, \epsilon_r - \delta_1) \psi_i^\mu \bar{\psi}_q^\nu \otimes P[r]_j^q w$$
$$= (\Lambda + \rho, \epsilon_r - \delta_1)(\omega + m - \overline{E})_i^k \bar{\psi}_{kq}^{\mu\nu}$$
$$\otimes P[r]_j^q w.$$

It follows that we may write, in terms of shift components,

$$\psi[l]_i^\mu \psi[r]_j^\nu \bar{\psi} \otimes w$$
$$= (\Lambda + \rho, \epsilon_r - \delta_1)(\omega + m - \bar{\alpha}_l) \bar{\psi}[l{:}r]_{ij}^{\mu\nu} w, \tag{28}$$

where we have employed the gl(m) adjoint identity (25) and where $\bar{\psi}[l{:}r]$ denote the "primary" shift components

$$\bar{\psi}[l{:}r]_{ij}^{\mu\nu} = \overline{P}[l]_i^k \bar{\psi}_{kq}^{\mu\nu} P[r]_j^q. \tag{29}$$

In this case $\overline{\alpha}_l = l - 1 - \lambda'_l$, where λ' now refers to the shifted gl(m) highest weight $\lambda' = \lambda - \epsilon_r - \epsilon_l$, so that

$$\lambda'_l = \begin{cases} \lambda_l - 1, & l \neq r, \\ \lambda_l - 2 & l = r. \end{cases}$$

Substituting into Eq. (28) we arrive at

$$\psi[l]_i^\mu \psi[r]_j^\nu \bar{\psi} \otimes w = A_{l,r} \bar{\psi}[l{:}r]_{ij}^{\mu\nu} w, \tag{30}$$

where the coefficients $A_{l,r}$, referred to herein as "selection factors," are given by

$$A_{l,r} = (\Lambda + \rho, \epsilon_r - \delta_1) \times \begin{cases} (\Lambda + \rho, \epsilon_l - \delta_1), & l \neq r, \\ (\Lambda + \rho, \epsilon_r - \delta_2), & l = r \end{cases}$$

and we have employed Eq. (26).

In the case $l = r$ above, the tensor $\bar{\psi}[r{:}r]_{ij}^{\mu\nu}$ is necessarily symmetric in i and j and hence antisymmetric in μ and ν. In such a case the states (30) are either identically zero or else span the irreducible gl(m) $\oplus$ gl(n) submodule of $V(\Lambda)$ with highest weight $\Lambda_0 = \Lambda - (\epsilon_r - \delta_1) - (\epsilon_r - \delta_2)$. In the case $l \neq r$ however, Eq. (30) does not yield a (pure shift) tensor and it is necessary to apply symmetrization operations to the superscripts μ, ν (symmetric and antisymmetric

in this case). We then obtain, after symmetrizing the shift labels (c.f. Ref. 24), the following pure shift tensor equations ($l \neq r$):

$$\psi_i^{[\mu}\psi_j^{\nu]}[l,r]\bar{\psi}\otimes w = A_{l,r}\bar{\psi}[l,r]_{ij}^{[\mu,\nu]}w, \tag{31a}$$

$$\psi_i^{\{\mu}\psi_j^{\nu\}}[l,r]\bar{\psi}\otimes w = A_{l,r}\bar{\psi}[l,r]_{ij}^{\{\mu,\nu\}}w, \tag{31b}$$

where

$$\psi_i^{[\mu}\psi_j^{\nu]} = (1/\sqrt{2})(\psi_i^{\mu}\psi_j^{\nu} - \psi_i^{\nu}\psi_j^{\mu}),$$

$$\bar{\psi}_{ij}^{[\mu,\nu]} = (1\sqrt{2})(\bar{\psi}_{ij}^{\mu\nu} - \bar{\psi}_{ij}^{\nu\mu})$$

are antisymmetric in μ and ν (and thus symmetric in i and j) whilst

$$\psi_i^{\{\mu}\psi_j^{\nu\}} = (1/\sqrt{2})(\psi_i^{\mu}\psi_j^{\nu} + \psi_i^{\nu}\psi_j^{\mu}),$$

$$\bar{\psi}_{ij}^{\{\mu,\nu\}} = (1/\sqrt{2})(\bar{\psi}_{ij}^{\mu\nu} + \bar{\psi}_{ij}^{\nu\mu})$$

are symmetric in μ and ν (and thus anti-symmetric in i and j). The $l \neq r$ shift components of the above tensors are then given by

$$\bar{\psi}[l,r]_{kj}^{[\mu,\nu]} = \bar{\psi}_{kq}^{[\mu,\nu]}P_S[l,r]_{ij}^{kq}$$

$$= \bar{\psi}[l{:}r]_{ij}^{[\mu,\nu]} + \bar{\psi}[r{:}l\,]_{ij}^{[\mu,\nu]},$$

$$\bar{\psi}[l,r]_{ij}^{\{\mu,\nu\}} = \bar{\psi}_{kq}^{\{\mu,\nu\}}P_A[l,r]_{ij}^{kq}$$

$$= \bar{\psi}[l{:}r]_{ij}^{\{\mu,\nu\}} + \bar{\psi}[r{:}l\,]_{ij}^{\{\mu,\nu\}},$$

where $P_S[l,r]$, $P_A[l,r]$ denote the appropriate $\mathrm{gl}(m)$ second rank symmetric and antisymmetric tensor projections, respectively.[34]

The states (31a) are either identically zero or else span the irreducible $\mathrm{gl}(m)\oplus\mathrm{gl}(n)$ submodule of $V(\Lambda)$ with highest weight $\Lambda_0 = \Lambda - (\epsilon_r - \delta_1) - (\epsilon_l - \delta_2)$, whilst the states (31b) are either zero or else span the irreducible $\mathrm{gl}(m)\oplus\mathrm{gl}(n)$ submodule with highest weight $\Lambda_0 = \Lambda - (\epsilon_r - \delta_1) - (\epsilon_l - \delta_1)$.

We write the above equations in the unified form

$$\psi_i^{\mu}\psi_j^{\nu}[l,r]\bar{\psi}\otimes w = A_{l,r}\bar{\psi}[l,r]_{ij}^{\mu\nu}w, \tag{32}$$

where

$$\bar{\psi}[l,r]_{ij}^{\mu\nu} = \bar{\psi}_{kq}^{\mu\nu}P[l,r]_{ij}^{kq}$$

with $P[l,r]$ the symmetric or antisymmetric second rank tensor projection of $\mathrm{gl}(m)$.[34] In the case that $P[l,r]$ is symmetric (resp. antisymmetric) the shift components in Eq. (32) are symmetric (resp. antisymmetric) in i and j and thus antisymmetric (resp. symmetric) in μ and ν: when $l = r$ the projection $P[r,r]$ is automatically symmetric[34] so that the tensor (32) is symmetric in i and j.

In general if we have a product of generators

$$\psi_{i_1}^{\mu_1}\cdots_{i_k}^{\mu_k} = \psi_{i_1}^{\mu_1}\cdots\psi_{i_k}^{\mu_k}$$

then we have a resolution into shift components

$$\psi[\Delta]_{i_1}^{\mu_1}\cdots_{i_k}^{\mu_k} = \psi_{j_1}^{\mu_1}\cdots_{j_k}^{\mu_k}P[\Delta]_{i_1}^{j_1}\cdots_{i_k}^{j_k},$$

where $P[\Delta]$ denotes a k th-rank irreducible tensor projection,[34] which effects the following shifts on the $\mathrm{gl}(m)\oplus\mathrm{gl}(n)$ representation labels:

$$\psi[\Delta]_{i_1}^{\mu_1}\cdots_{i_k}^{\mu_k}v\in V_0(\lambda - \Delta|\gamma + (\dot{\omega})),\quad v\in V_0(\Lambda). \tag{33}$$

Here γ is the highest weight of an irreducible k th-rank tensor representation of $\mathrm{gl}(n)$, whose corresponding Young diagram $[\gamma]$ determines the symmetry on the indices $\mu_1,\cdots\mu_k$, and where Δ is necessarily a $\mathrm{gl}(m)$ weight in the irreducible $\mathrm{gl}(m)$ module given by the conjugate Young diagram $[\gamma]^T$. In particular, since the Young diagram $[\gamma]^T$ cannot have more than n columns, the components of the shift weight Δ must satisfy

$$0\leqslant\Delta_r\leqslant n,(1\leqslant r\leqslant m),\quad \sum_{r=1}^{m}\Delta_r = k. \tag{34}$$

Our aim now is to prove, by induction, that

$$\psi[\Delta]_{i_1\cdots i_k}^{\mu_1\cdots\mu_k}\bar{\psi}\otimes w = A_\Delta\bar{\psi}[\Delta]_{i_1\cdots i_k}^{\mu_1\cdots\mu_k}w$$

$$= A_\Delta\bar{\psi}_{j_1\cdots j_k}^{\mu_1\cdots\mu_k}\otimes P[\Delta]_{i_1\cdots i_k}^{j_1\cdots j_k}w \tag{35}$$

where the selection factors A_Δ are given by the following product of "atypicality factors":

$$A_\Delta = \prod_{r=1}^{m}\prod_{\mu=1}^{\Delta_r}(\Lambda + \rho,\epsilon_r - \delta_\mu). \tag{36}$$

In view of Eq. (34) the products in Eq. (36) are always well defined.

To prove Eq. (35) inductively we have, using Eq. (22b),

$$\psi_i^{\mu}\psi[\Delta]_{i_1\cdots i_k}^{\mu_1\cdots\mu_k}\bar{\psi}\otimes w = A_\Delta\psi_i^{\mu}\bar{\psi}_{j_1\cdots j_k}^{\mu_1\cdots\mu_k}\otimes P[\Delta]_{i_1\cdots i_k}^{j_1\cdots j_k}w$$

$$= A_\Delta(\omega + m - \overline{E})_i^q\bar{\psi}_{qj_1\cdots j_k}^{\mu\mu_1\cdots\mu_k}$$

$$\otimes P[\Delta]_{i_1\cdots i_k}^{j_1\cdots j_k}w.$$

Applying the $\mathrm{gl}(m)$ adjoint projector $\overline{P}[r]$ to the left we obtain

$$\psi[r]_i^{\mu}\psi[\Delta]_{i_1\cdots i_k}^{\mu_1\cdots\mu_k}\bar{\psi}\otimes w$$

$$= A_\Delta(\omega + m - \overline{\alpha}_r)\bar{\psi}[r{:}\Delta]_{ii_1}^{\mu\mu_1}\cdots_{i_k}^{\mu_k}w \tag{37}$$

where $\bar{\psi}[r{:}\Delta]$ denote the primary shift components

$$\bar{\psi}[r{:}\Delta]_{ii_1\cdots i_k}^{\mu\mu_1\cdots\mu_k} = \overline{P}[r]_i^q\bar{\psi}_{qj_1\cdots j_k}^{\mu\mu_1\cdots\mu_k}P[\Delta]_{i_1\cdots i_k}^{j_1\cdots j_k}.$$

In this case [cf. Eq. (23)] we have $\overline{\alpha}_r = r - 1 - \lambda'_r$, where λ' now refers to the shifted $\mathrm{gl}(m)$ highest weight $\lambda' = \lambda - \Delta - \epsilon_r$. We may thus write, in view of Eq. (26),

$$\omega + m - \overline{\alpha}_r = \omega + m - r + \Delta_r + \lambda_r$$

$$= (\Lambda + \rho,\epsilon_r - \delta_{\Delta_r+1})$$

and substituting into Eq. (37) we arrive at

$$\psi[r]_i^{\mu}\psi[\Delta]_{i_1}^{\mu_1}\cdots_{i_k}^{\mu_k}\bar{\psi}\otimes w = A_{\Delta+\epsilon_r}\bar{\psi}[r{:}\Delta]_{ii_1}^{\mu\mu_1}\cdots_{i_k}^{\mu_k}w.$$

Finally, by symmetrizing the $\mathrm{gl}(m)$ shift labels,[28] we arrive at the pure shift tensor equation

$$\psi[\Delta+\epsilon_r]_{ii_1}^{\mu\mu_1}\cdots_{i_k}^{\mu_k}\bar{\psi}\otimes w = A_{\Delta+\epsilon_r}\bar{\psi}[\Delta+\epsilon_r]_{ii_1}^{\mu\mu_1}\cdots_{i_k}^{\mu_k}w,$$

which establishes Eq. (35) by induction.

The states (35) [or equivalently (33)] are either identically zero, which will occur if $A_\Delta = 0$ or

$$\Lambda_0 = (\lambda - \Delta|(\dot{\omega}) + \gamma) \tag{38}$$

is nondominant, or else they span a (possible multiple of the) irreducible $\mathrm{gl}(m)\oplus\mathrm{gl}(n)$ submodule of $V(\Lambda)$ with highest weight Λ_0. It follows that the irreducible $\mathrm{gl}(m)\oplus\mathrm{gl}(n)$ modules occurring in $V(\Lambda)$ have highest

weights of the form (38), where γ is the highest weight of an irreducible k th-rank tensor representation of $\mathrm{gl}(n)$ and Δ is a $\mathrm{gl}(m)$ (shift) weight in the $\mathrm{gl}(m)$ irrep. with highest weight $\bar{\gamma}$ whose Young diagram $[\bar{\gamma}] = [\gamma]^T$ is conjugate to $[\gamma]$. The irreducible $\mathrm{gl}(m) \oplus \mathrm{gl}(n)$ module $V_0(\Lambda_0)$ then occurs in $V(\Lambda)$ provided Λ_0 is dominant and $A_\Delta \neq 0$. In such a case the irreducible module $V_0(\Lambda_0)$ may occur with multiplicities (possibly zero) in $V(\Lambda)$. Indeed if Λ_0 is dominant and $A_\Delta \neq 0$, then the multiplicity of $V_0(\Lambda_0)$ in $V(\Lambda)$ is given by

$$m_{\Lambda_0} = m_{\lambda - \Delta} m_\gamma,$$

where $m_{\lambda-\Delta}$ is the multiplicity of the irreducible $\mathrm{gl}(m)$ module $V_0(\lambda - \Delta)$ in the $\mathrm{gl}(m)$ tensor product

$$V_0(\lambda) \otimes V_0^*(\bar{\gamma})$$

and m_γ is the number of irreducible k th-rank $\mathrm{gl}(n)$ tensors of symmetry type γ, which may be evaluated using standard Young diagram techniques.[35]

In this way we may deduce the $\mathrm{gl}(m|n) \downarrow \mathrm{gl}(m) \oplus \mathrm{gl}(n)$ branching rules (and thus characters) for the irreps of $\mathrm{gl}(m|n)$ with highest weights of the special form $\Lambda = (\lambda \,|\dot{\omega})$. However we shall not pursue this any further here since it turns out that a more efficient procedure can be employed based on the Gel'fand–Tsetlin (GT) chain.

IV. gl(m|n)↓gl(m|n−1) BRANCHING RULES

Perhaps the most surprising thing about the class of irreps $V(\lambda\,|\dot{\omega})$ we are considering is that they decompose into a direct sum of irreducible representations of $\mathrm{gl}(m|n-1)$ of the same type. To see this we apply the PBW theorem and order the odd lowering generators ψ_i^μ with the operators ψ_i^n $(1 \leqslant i < m)$ appearing on the right. With this ordering the operator $\bar{\psi}$ of Eq. (15) may be decomposed

$$\bar{\psi} = \bar{\psi}_0 \tilde{\psi} \tag{39}$$

with

$$\tilde{\psi} = \prod_{i=1}^{m} \psi_n^i$$

and where $\bar{\psi}_0$ is the $\mathrm{gl}(m|n-1)$ analog of $\bar{\psi}$:

$$\bar{\psi}_0 = \prod_{i=1}^{m} \prod_{\mu=1}^{n-1} \psi_\mu^i.$$

To write the tensors of Eq. (17) in $\mathrm{gl}(m) \oplus \mathrm{gl}(n-1)$ symmetry adapted form, we also find it convenient to introduce the antisymmetric $\mathrm{gl}(m)$ tensors $\tilde{\psi}_{i_1 \cdots i_k}$ defined recursively according to

$$\begin{aligned}
&\psi^j{}_n \tilde{\psi}_i = \delta^j{}_i \tilde{\psi} \\
&\psi^j{}_n \tilde{\psi}_{i_1 \cdots i_k} = \delta^j{}_{i_1} \tilde{\psi}_{i_2 \cdots i_k} - \delta^j{}_{i_2} \tilde{\psi}_{i_1 i_3 \cdots i_k} \\
&\qquad + \cdots + (-1)^{k-1} \delta^j{}_{i_k} \tilde{\psi}_{i_1 \cdots i_{k-1}}.
\end{aligned} \tag{40}$$

With the above ordering the irreducible $\mathrm{gl}(m|n)$ module $V(\Lambda)$, $\Lambda = (\lambda\,|\dot{\omega})$, is generated by applying all odd lowering $\mathrm{gl}(m|n-1)$ generators to states of the form

$$\psi_{i_1}^n \cdots \psi_{i_k}^n \bar{\psi} \otimes w, \quad w \in V_0(\Lambda - 2\rho_1). \tag{41}$$

In other words $V(\Lambda)$ is given by the $\mathrm{gl}(m|n-1)$ module generated by the states (41). We note, since the $\mathrm{gl}(n)$ component of the highest weight Λ is trivial, that the irreducible $\mathrm{gl}(m) \oplus \mathrm{gl}(n)$ module $V_0(\Lambda - 2\rho_1)$ is also irreducible for the canonical subalgebra $\mathrm{gl}(m) \oplus \mathrm{gl}(n-1)$.

On the other hand, we note that

$$\psi_{i_1}^n \cdots \psi_{i_k}^n$$

constitutes a contragradient antisymmetric tensor operator of $\mathrm{gl}(m)$ and hence we have the following resolution into antisymmetric shift components:

$$\psi_{i_1}^n \psi_{i_k}^n = \sum_\Delta \psi[\Delta]^n_{i_1} \cdots {}^n_{i_k}$$

where the sum is over all weights

$$\Delta = \epsilon_{r_1} + \epsilon_{r_2} + \cdots + \epsilon_{r_k}, 1 \leqslant r_1 < r_2 < \cdots < r_k \leqslant m$$

in the anti-symmetric tensor representation of $\mathrm{gl}(m)$ and $\psi[\Delta]$ effects the shift

$$\lambda \backsim \lambda - (\epsilon_{r_1} + \cdots + \epsilon_{r_k})$$

in the $\mathrm{gl}(m)$ representation labels. The above shift components are given explicitly by

$$\psi[\Delta]^n_{i_1} \cdots {}^n_{i_k} = \psi_{j_1}^n \cdots \psi_{j_k}^n j_k P[\Delta]^{j_1 \cdots j_k}_{i_1 \cdots i_k},$$

where $P[\Delta]$ is the $\mathrm{gl}(m)$ k th-rank antisymmetric tensor projection[34] as applied in Ref. 24.

The vectors of Eq. (41) may thus be resolved into $\mathrm{gl}(m) \oplus \mathrm{gl}(n-1)$ symmetry adapted components

$$\psi_{i_1}^n \cdots \psi_{i_k}^n \bar{\psi} \otimes w = \sum_\Delta \psi[\Delta]^n_{i_1} \cdots {}^n_{i_k} \bar{\psi} \otimes w \tag{42}$$

with

$$\psi[\Delta]^n_{i_1} \cdots {}^n_{i_k} \bar{\psi} \otimes w = A_\Delta \bar{\psi}^{n \cdots n}_{j_1 \cdots j_k} \otimes P[\Delta]^{j_1 \cdots j_k}_{i_1 \cdots i_k} w,$$

where in the above we have employed Eq. (35) and the selection factor A_Δ is given by [cf. Eq. (36)]

$$A_\Delta = \prod_{i=1}^{k} (\Lambda + \rho, \epsilon_{r_i} - \delta_1).$$

In terms of the tensors (40) the above action may alternatively be written

$$\begin{aligned}
\psi[\Delta]^n_{i_1} \cdots {}^n_{i_k} \bar{\psi} \otimes w &= A_\Delta \bar{\psi}_0 \tilde{\psi}_{j_1 \cdots j_k} \otimes P[\Delta]^{j_1 \cdots j_k}_{i_1 \cdots i_k} w \\
&= A_\Delta \bar{\psi}_0 \otimes w_\Delta,
\end{aligned} \tag{43}$$

where

$$w_\Delta = \tilde{\psi}_{j_1 \cdots j_k} \otimes P[\Delta]^{j_1 \cdots j_k}_{i_1 \cdots i_k} w. \tag{44}$$

In view of the special nature of antisymmetric tensors (c.f. Ref. 36, Appendix A) we note that the states (43) are either identically zero, for all $w \in V_0(\Lambda - 2\rho_1)$ and choice of indices $1 \leqslant i_1, i_2, \ldots, i_k \leqslant m$, or else they necessarily span an *irreducible* $\mathrm{gl}(m) \oplus \mathrm{gl}(n-1)$ module with highest weight $\Lambda_0 = (\lambda - \Delta\,|\dot{\omega})$. In view of Eq. (43) the latter case arises if and only if Λ_0 is dominant and $A_\Delta \neq 0$.

We now claim that the states (43), if not zero, in fact generate an irreducible $\mathrm{gl}(m|n-1)$ module with highest weight $(\lambda - \Delta\,|\dot{\omega})$. To see this we note that the vectors ω_Δ of Eq. (44) must satisfy, for $1 \leqslant i \leqslant m$, $1 \leqslant \mu < n$

$$\psi_i^\mu w_\Delta = \psi_i^\mu \tilde{\psi}_{j_1 \cdots j_k} \otimes P[\Delta]^{j_1 \cdots j_k}_{i_1 \cdots i_k} w = 0,$$

which follows because w belongs to a trivial one-dimensional

$\mathrm{gl}(n)$ module (so that $E_n^\mu w = 0$, $\mu < n$). Thus all $\mathrm{gl}(m|n-1)$ odd lowering operators ψ_i^μ vanish on the states (44) so that, from the modified induced module construction of Eq. (14), the $\mathrm{gl}(m|n-1)$ module generated by the states (43) is necessarily irreducible with $\mathrm{gl}(m|n-1)$ highest weight $(\lambda - \Delta|\dot{\omega})$.

In view of the resolution (42) is follows that the irreducible $\mathrm{gl}(m|n)$ module $V(\Lambda)$, $\Lambda = (\lambda|\dot{\omega})$, is a direct sum of irreducible $\mathrm{gl}(m|n-1)$ modules with highest weights

$$(\lambda - \Delta|\dot{\omega}), \quad \Delta = \epsilon_{r_1} + \cdots + \epsilon_{r_k},$$
$$1 \leqslant r_1 < r_2 < \cdots < r_k \leqslant m$$

subject to the condition that $\lambda - \Delta$ is $\mathrm{gl}(m)$-dominant and

$$\prod_{i=1}^{k} (\Lambda + \rho, \epsilon_{r_i} - \delta_1) \neq 0,$$

each such module occurring exactly once. We thus arrive at the following result.

Theorem 2: The highest weights of the irreducible $\mathrm{gl}(m|n-1)$ modules occurring in the irreducible $\mathrm{gl}(m|n)$ module $V(\Lambda)$, $\Lambda = (\lambda|\dot{\omega})$, are of the form $(\lambda_0|\dot{\omega})$, where λ_0 is a dominant weight of $\mathrm{gl}(m)$ satisfying $\lambda_i - \lambda_{0_j} \in \mathbb{Z}$ together with

$$\lambda_i \geqslant \lambda_{0_i} \geqslant \lambda_i - 1, \quad (\Lambda + \rho, \epsilon_i - \delta_1) \neq 0,$$
$$\lambda_{0_i} = \lambda_i, \quad (\Lambda + \rho, \epsilon_i - \delta_1) = 0,$$

each such module occurring exactly once. ∎

We remark that since Λ is $\mathrm{gl}(m) \oplus \mathrm{gl}(n)$ dominant we can have $(\Lambda + \rho, \epsilon_i - \delta_1) = 0$ for at most one index i: note, in view of Eq. (26), that

$$(\Lambda + \rho, \epsilon_i - \delta_1) = \lambda_i + \omega + m - i.$$

We note further that if $(\Lambda + \rho, \epsilon_i - \delta_1) = 0$ for some index i then we must have for the $\mathrm{gl}(m|n-1)$ highest weights $(\lambda_0|\dot{\omega})$ in Theorem 2,

$$\lambda_{o_j} = \lambda_j$$

for any index $j < i$ such that $(\Lambda, \epsilon_j - \epsilon_i) = 0$. This possibility is automatically taken into account in Theorem (2) by requiring that λ_0 be $\mathrm{gl}(m)$ dominant. In the case $n = 1$ the results of the above theorem agree with the $\mathrm{gl}(m|1) \downarrow \mathrm{gl}(m)$ branching rules previously obtained by Gould, Bracken, and Hughes.[28]

V. THE GT BASIS

The results of the previous section demonstrate that the irreducible $\mathrm{gl}(m|n)$ module $V(\Lambda)$, with highest weight of the special form $\Lambda = (\lambda|\dot{\omega})$, is necessarily a direct sum of irreducible modules for the canonical subalgebra $\mathrm{gl}(m|n-1)$ with highest weights $(\lambda_0|\dot{\omega})$ of the same type. By repeated application of Theorem 2 it follows, from standard arguments, that we may choose a basis for $V(\Lambda)$ which is symmetry adapted to the canonical subalgebra chain

$$\mathrm{gl}(m|n) \supset \mathrm{gl}(m|n-1) \supset \cdots \supset \mathrm{gl}(m|1)$$
$$\supset \mathrm{gl}(m) \supset \mathrm{gl}(m-1) \supset \cdots \supset \mathrm{gl}(1). \tag{45}$$

The resulting GT basis states may be written

$$\left| \begin{matrix} (\Lambda) \\ (\lambda) \end{matrix} \right\rangle_\omega, \tag{46a}$$

where (Λ) denotes the pattern

$$(\Lambda) = \begin{pmatrix} \lambda_{1,m+n} & \cdots & \lambda_{m,m+n} \\ \lambda_{1,m+n-1} & \cdots & \lambda_{m,m+n-1} \\ \vdots & & \vdots \\ \lambda_{1,m+1} & \cdots & \lambda_{m,m+1} \end{pmatrix} \tag{46b}$$

and (λ) refers to a familiar $\mathrm{gl}(m)$ GT pattern

$$(\lambda) = \begin{pmatrix} \lambda_{1,m} & \cdots & \lambda_{m,m} \\ \lambda_{1,m-1} & \cdots & \lambda_{m-1,m-1} \\ & \iddots & \iddots \\ & \lambda_{1,2} \quad \lambda_{2,2} & \\ & \lambda_{1,1} & \end{pmatrix}. \tag{46c}$$

The entries $\lambda_{i,m+k}$ of the upper pattern (Λ), corresponding to the rep. labels $(\lambda_{m+k}|\dot{\omega})$ of the canonical subalgebra $\mathrm{gl}(m|k)$ $(0 < k < n)$, must satisfy $\lambda_{i,j} - \lambda_{k,l} \in \mathbb{Z}$ together with the betweeness conditions $(1 \leqslant i \leqslant m)$

$$\lambda_{i,m+k+1} \geqslant \lambda_{i,m+k} \geqslant \lambda_{i,m+k+1} - 1,$$
$$\omega + m - i + \lambda_{i,m+k+1} \neq 0;$$
$$\lambda_{i,m+k} = \lambda_{i,m+k+1}, \ \omega + m - i + \lambda_{i,m+k+1} = 0,$$

with

$$\lambda_{i,m+k} \geqslant \lambda_{i+1,m+k}, \quad 1 \leqslant i < m,$$

which follows from the branching rule of Theorem 2. As usual the top row $\lambda_{m+n} = (\lambda_{1,m+n}, \ldots, \lambda_{m,m+n})$ of the pattern (Λ) determines the highest weight $\Lambda = (\lambda_{m+n}|\dot{\omega})$ of the irreducible $\mathrm{gl}(m|n)$ module concerned.

We note that the GT states (46a) are uniquely determined up to scalar multiples and are eigenstates of the Casimir invariants for each of the canonical subalgebras in the chain (45). In particular the state (46a) is an eigenstate of the first order invariant $I_{1,m+k}$ of the canonical subalgebra $\mathrm{gl}(m|k)$, $0 \leqslant k \leqslant n$, with eigenvalue

$$k\omega + \sum_{i=1}^{m} \lambda_{i,m+k}.$$

It follows that the state (46a) is a $\mathrm{gl}(m|n)$ weight state of weight $\Lambda' = (\nu|\mu)$, where

$$\mu_k = \omega + \sum_{i=1}^{m} (\lambda_{i,m+k} - \lambda_{i,m+k-1}), \quad 1 \leqslant k \leqslant n$$

and, as usual,

$$\nu_j = \sum_{i=1}^{j} \lambda_{i,j} - \sum_{i=1}^{j-1} \lambda_{i,j-1}, \quad 1 \leqslant j \leqslant m.$$

We may clearly choose an inner product on $V(\Lambda)$ in which the above GT states are orthonormal. However such a choice of inner produce may not necessarily be most convenient. An alternative procedure is to adopt the approach of Ref. 29 and choose a symmetric non-degenerate sequilinear form $(,)$ on $V(\Lambda)$ under which we have the useful Hermiticity conditions

$$(E_j^i v,w) = (v,E_i^j w), \quad (E_\nu^\mu v,w) = (v,E_\mu^\nu w),$$
$$(\psi_\mu^i v,w) = (v,\psi_i^\mu w), \quad \text{for all } v,w \in V(\Lambda).$$

Such a generalized inner product is not necessarily positive definite but nevertheless the GT states (46a) are orthogonal under the form (,). Moreover the form (,) when restricted to all GT patterns with a fixed upper pattern (Λ), corresponding to a given irreducible $\mathrm{gl}(m)$ submodule, will either be positive or negative definite. If the form (,) is positive definite on all of $V(\Lambda)$, in which case it gives rise to an inner product, then $V(\Lambda)$ is necessarily a type (1) *-representation of $\mathrm{gl}(m|n)$.[29]

It would be of interest to determine the matrix elements of the $\mathrm{gl}(m|n)$ generators in the GT basis (46), under the above mentioned form (,), and in particular to examine in detail the *-representations of $\mathrm{gl}(m|n)$ with highest weights of the form $(\lambda\,|\dot{\omega})$, since these are likely to be of great physical interest (e.g., the interacting boson-fermion model of the nucleus); those highest weights corresponding to *-irreps can be obtained directly from the classification scheme of Gould and Zhang.[37]

It would also be of interest to determine the lowest weights and characters of the irreducible $\mathrm{gl}(m|n)$ modules $V(\lambda\,|\dot{\omega})$. This can be done in the special case that $V(\lambda\,|\dot{\omega})$ corresponds to a tensor irrep.[13,15,38] of $\mathrm{gl}(m|n)$, in which case ω is necessarily real. From the results of Gould and Zhang,[37] $V(\lambda\,|\dot{\omega})$ corresponds to a contravariant tensor irrep. when $(\lambda_m + \omega) \in \mathbb{Z}^+$ and to a covariant tensor irrep. when one of the following conditions are satisfied

$$\lambda_1 + \omega = -k, \quad m \leqslant k \in \mathbb{Z}^+$$

or

$$\lambda_1 + \omega = -k, \quad (\lambda, \epsilon_1 - \epsilon_{m-k}) = 0, \quad 0 \leqslant k < m.$$

For this class of irreps, and those obtained from them by tensor products with a one-dimensional irrep., the lowest weights are given explicitly by Gould and Zhang[37] and the corresponding characters by the Serganova formula[39,40] and it would clearly be of interest to determine analogous results for the remaining irreducible $\mathrm{gl}(m|n)$ modules $V(\lambda\,|\dot{\omega})$. In particular it would be of interest to determine which of the known character formulae,[10,22,23,39,40] if any, hold for the above class of irreducible $\mathrm{gl}(m|n)$ modules.

[1]Y. Ne'eman and S. Sternberg, Proc. Natl. Acad. Sci. (USA) **77**, 3127 (1980).
[2]V. A. Kostelecky and D. K. Campbell, Physica D **15**, 3 (1985).
[3]G. Parisi and N. Sourlas, Phys. Rev. Lett. **43**, 744 (1979).
[4]Y. Nambu, Physica D **15**, 147 (1985).
[5]A. Montorsi, M. Rasetti, and A. I. Solomon, Phys. Rev. Lett. **59**, 2243 (1987).
[6]F. Iachello, Phys. Rev. Lett. **44**, 772 (1980).
[7]J. Wess and B. Zumino, Nucl. Phys. B **70**, 39 (1974).
[8]V. G. Kac, Adv. Math. **26**, 8 (1977).
[9]V. G. Kac, *Lecture Notes in Mathematics*, Vol. 676 (Springer-Verlag, Berlin, 1978), p. 597.
[10]I. N. Bernstein and D. A. Leites, C. R. Acad. Bulg, Sci. **33**, 1049 (1980).
[11]H. Sun and Q. Han, Phys. Energia Fortis Phys. Nucl. **5**, 546 (1981).
[12]J. Thierry-Mieg and B. Morel, in *Superspace and Supergravity*, edited by S. Hawking and M. Rocek (Cambridge U.P., Cambridge, 1981).
[13]P. H. Dondi and P. D. Jarvis, J. Phys. A: Math. Gen **14**, 547 (1981).
[14]J. W. B. Hughes, J. Math. Phys. **22**, 245 (1981).
[15]A. B. Balantekin and I. Bars, J. Math. Phys. **22**, 1810 (1981); **23**, 1239 (1982).
[16]J. P. Hurni and B. Morel, J. Math. Phys. **22**, 2236 (1982); **24**, 157 (1983).
[17]R. C. King, *Lecture Notes in Physics*, Vol. 180 (Springer-Verlag, Berlin, 1983), p. 41.
[18]R. J. Farmer and P. D. Jarvis, J. Phys. A: Math. Gen. **16**, 473 (1983); **17**, 2365 (1984).
[19]J. Van der Jeugt, J. Math. Phys. **25**, 3334 (1984).
[20]J. Thierry-Mieg, Phys. Lett. B **138**, 393 (1984).
[21]F. Deluc and M. Gourdin, J. Math. Phys. **25**, 1651 (1984).
[22]J. W. B. Hughes and R. C. King, J. Phys. A: Math. Gen. **20**, L1047 (1987).
[23]J. Van der Jeugt, J. Phys. A: Math. Gen. **20**, 209 (1987).
[24]J. P. Hurni, J. Phys. A: Math. Gen. **20**, 5755 (1987).
[25]T. D.Palev, J. Math. Phys. **28**, 2280 (1987); **29**, 2589 (1988).
[26]T. D. Palev, Func. Anal. Appl. **21**, 245 (1987); see also "Irreducible finite-dimensional representations of the Lie suberalgebra $\mathrm{gl}(n|1)$ in a Gel'fand–Zetlin basis," Preprint, IC/88/207 (Trieste, 1988), to appear in J. Math. Phys.
[27]M. D. Gould, J. Phys. A: Math. Gen. **22**, 1209 (1989).
[28]M. D. Gould, A. J. Bracken, and J. W. B. Hughes, J. Phys. A: Math. Gen. **22**, 2879 (1989).
[29]M. D. Gould and R. B. Zhang, J. Math. Phys. **31**, 1524 (1990).
[30]T. D. Palev, "Essentially typical representations of the Lie superalgebra $\mathrm{gl}(n,m)$ in a Gel'fand-Tsetlin basis," (in Russian) Preprint, JINR P5-88-169 (Dubna, 1988).
[31]M. Scheunert, *Lecture Notes in Mathematics*, Vol. **716** (Springer-Verlag, Berlin, 1979).
[32]J. Vervier, Riv. Nuovo Cimento **10**, 1 (1987).
[33]A. J. Bracken and H. S. Green. J. Math. Phys. **12**, 2099 (1971).
[34]H. S. Green, J. Math. Phys. **12**, 2106 (1971); J. Austral. Math. Soc. **B19**, 129 (1975).
[35]See for example M. Hamermesh, *Group Theory* (Addison-Wesley, Reading, MA, 1962); B. G. Wybourne *Classical Groups for Physicists* (Wiley, New York, 1974).
[36]M. D. Gould, Rep. Math. Phys. **27**, 73 (1989).
[37]M. D. Gould and R. B. Zhang, J. Math. Phys. (to appear); Lett. Math. Phys. (to appear).
[38]A. Berele and A. Regev, Adv. Math. **64**, 118 (1987).
[39]V. Serganova and A. Serge'ev, "Super Weyl groups and dominant weights," Dept. of Math. Rep. Univ. of Stockholm, Sweden (1987).
[40]J. Van der Jeugt, J. W. B. Hughes, R.C. King, and J. Thierry-Mieg, "Character formulas for irreducible modules of the Lie superalgebras $\mathrm{sl}(m|n)$," Preprint, Queen Mary College, London (1989).

Local realizations of kinematical groups with a constant electromagnetic field. II. The nonrelativistic case

Javier Negro and Mariano A. del Olmo
Departamento de Física Teórica, Facultad de Ciencias, Universidad de Valladolid. 47011 Valladolid, Spain

(Received 14 December 1989; accepted for publication 5 July 1990)

In this paper, nonrelativistic elementary physical systems interacting with constant external electromagnetic fields are studied. The method is to construct a special kind of realizations of the Galilei group, which depend on the electromagnetic field. The linearization of this problem, which consists in obtaining these local realizations via the linear representations of another group, leads to a new representation group: the nonrelativistic Maxwell group. The study of the representations of this group and the related invariant equations completes this work.

I. INTRODUCTION

In a preceding paper,[1] hereafter called I, we started the study, in a group-theoretical context, of elementary physical systems interacting with external constant electromagnetic fields (emf). That work dealt with relativistic systems, and here we present the nonrelativistic case.

The origin of this group-theoretical treatment is the well-known paper by Wigner[2] of 1939. There, the quantum elementary systems were put in correspondence with the semiunitary projective irreducible representations of the symmetry group of the physical system. Later, Hoogland[3–5] showed that the interesting representations from a physical point of view in the space-time description were those that he characterized as locally operating representations (lor). The mathematical development of these representations can be found in several papers.[6–10] These local representations allow us to obtain the differential equations that describe the free systems, but such a method is not so straightforward when one tries to study interacting systems because there does not exist a general formulation.

In I we gave a group-theoretical description for the case of a constant emf, defining a new kind of local realizations of the Poincaré group P, which also depends on the emf f by

$$(U(g)\psi)(g(x,f))=A(g;x,f)\psi(x,f), \tag{1.1}$$

where A is a Borel nonsingular matrix-valued function: the gauge matrix. There exists a group such that these realizations can be obtained from some of its linear representations. This new group, minimal in some sense, which appears in this way is called the representation group,[8,11] and in this case, it is known as the Maxwell group.[12] The local representations of the Maxwell group are, in general, reducible, so to turn them into irreducible it is necessary to impose some extra conditions. These extra conditions will be written in the form of differential equations and characterize the interacting system.

In the present work we extend this formulation to the Galilei group, although the theory is also valid for any other kinematical group. The first problem arises when one tries to write the emf that supports this kinematical group. In our case we have followed the paper by Le Bellac and Lévy-Leblond[13] where they make an exhaustive study of the Galilean electromagnetism. The results we are presenting here are very close to the ones found in I, except that here there are two Galilean Maxwell groups, each of them related to the two limits of the Galilean electromagnetism, respectively.

Finally, we wish to note that our two papers, I and this one, complete and participate the ideas of former attempts[12–16] made to give a group-theoretical formulation of the electromagnetic interaction with elementary systems. It must be remarked that Beckers and Hussin,[16] making use of the Schrader's method, obtained a nonrelativistic Maxwell group which also can be found in our derivation.

The organization of this paper is as follows. In Sec. II we present a short review on the Galilei group, some of its finite-dimensional representations and the Galilean electromagnetism. Section III presents the local realizations of the Galilei group and the construction of the nonrelativistic Maxwell groups. In Sec. IV we study exhaustively the irreducible representations of the Maxwell groups and obtain the Casimirs that label these irreducible representations. We have computed the local representations of the Maxwell groups in Sec. V. In the next section we have derived some invariant equations for charged particles with spin 0 and $\frac{1}{2}$ linked with certain local representations of the Maxwell group. The electromagnetic minimal coupling appears in some of these representations in a natural way. Finally, we present some conclusions and remarks in Sec. VII.

II. GALILEAN ELECTROMAGNETIC FIELDS

The format of this section is in the nature of a review. First of all, we present some features about the Galilei group and study some of its finite representations which will be used later. The second part is devoted to the Galilean electromagnetism.

 0022-2488/90/122811-11$03.00

A. The Galilei group

The Galilei group G is the transformation group of the Newtonian space-time. It is a ten parameter Lie group whose composition law is

$$gg' \equiv (b',\mathbf{a}',\mathbf{v}',R')(b,\mathbf{a},\mathbf{v},R)$$
$$= (b'+b,\mathbf{a}'+R'\mathbf{a}+\mathbf{v}'b,\mathbf{v}'+R'\mathbf{v},R'R), \tag{2.1}$$

where $\mathbf{a}$, b, and $\mathbf{v}$ denote the parameters of the space-time translations, and pure Galilean transformations, respectively; and R is a generic SO(3) rotation. Sometimes we will write the elements $g \equiv (b,\mathbf{a},\mathbf{v},R)$ in the form $g \equiv (a,\Lambda)$, where a stands for $(b,\mathbf{a}) \in \mathbf{R}^4$ and Λ for $(\mathbf{v},R) \in G_0$ (the homogeneous Galilei group). The Galilei Lie algebra is defined by the following nonzero commutators:

$$[J_i,J_j]=\epsilon_{ijk}J_k, \quad [J_i,K_j]=\epsilon_{ijk}K_k,$$
$$[J_i,P_j]=\epsilon_{ijk}P_k, \quad [K_i,H]=P_i, \tag{2.1'}$$

where $i, j, k = 1, 2, 3$. The generators J_i and K_i correspond to the rotations around and inertial transformations across the i axis, respectively, and $\mathbf{P}$ and H to the space-time translations.

The second cohomology group of G with values in U(1) is $H^2(G,\mathrm{U}(1)) = \mathbf{R} \otimes Z_2$. Its connected part is associated with the central extensions of the Galilei Lie algebra given by the above nonzero commutators and the new ones $[K_i,P_j] = m\delta_{ij}I$, I being a new generator that commutes with any generator of the Galilei Lie algebra. The cohomology classes are labeled by $[m,l]$, $m \in \mathbf{R}$ and $l \in Z_2$. A lifting of a generic class $[m,l]$ is given by

$$\omega_{m,l}((b',\mathbf{a}',\mathbf{v}',R'),(b,\mathbf{a},\mathbf{v},R))$$
$$= \omega^{\mathrm{SO}(3)}{}_l(R',R)\exp\{im(\tfrac{1}{2}b\mathbf{v}'^2 + \mathbf{v}'\cdot R'\mathbf{a})\} \tag{2.2}$$

where $\omega^{\mathrm{SO}(3)}{}_l$ is a factor system of SO(3). This factor system lifts the class $[l]$ which belongs to $H^2(\mathrm{SO}(3),\mathrm{U}(1))$ and is defined by

$$\omega^{\mathrm{SO}(3)}{}_l(R',R)=1,$$
$$\omega^{\mathrm{SO}(3)}{}_{-l}(R',R)=\sigma(R')\sigma(R)\sigma^{-1}(R'R), \tag{2.3}$$

where σ is a normalized Borel section from SO(3) to SU(2). In the following, for the sake of simplicity, we will work with the universal covering group G^* of G. So we put SU(2) instead of SO(3) and define the action of SU(2) on the space-time through the canonical homomorphism from SU(2) to SO(3). In doing so we do not need to bother anymore for the extension given by $\omega^{\mathrm{SO}(3)}{}_l(R,R')$. However, we will use the same notation for the elements of G^* as well as those of G without risk of confusion.

The homogeneous Galilei group G_0 is a semidirect product $V \odot \mathrm{SO}(3)$ [now $V \odot \mathrm{SU}(2)$], where $V \approx \mathbf{R}^3$ is the subgroup of the Galilean inertial transformations. Two interesting finite-dimensional representations of G_0 are Δ_+ and D_-. The first one is supported by the points of the space-time:

$$\Delta_+(\mathbf{v},R):(t,\mathbf{x}) \to (t,R\mathbf{x}+\mathbf{v}t), \tag{2.4}$$

while the derivation operators (∂_t,∇) transform under D_-:

$$D_-(\mathbf{v},R):(\partial_t,\nabla) \to (\partial_t - \mathbf{v}\cdot R\nabla, R\nabla). \tag{2.5}$$

In general, we can define the representations Δ_λ and D_λ, $\lambda \in \mathbf{R}^*$ by

$$\Delta_\lambda(\mathbf{v},R):(b,\mathbf{a}) \to (b,R\mathbf{a}+\lambda\mathbf{v}t),$$
$$D_\lambda(\mathbf{v},R):(\beta,\alpha) \to (\beta+\lambda\mathbf{v}\cdot R\alpha, R\alpha). \tag{2.6}$$

The representations Δ_λ are equivalent, and so are the D_λ. However, no Δ_λ is equivalent to any $D_{\lambda'}$.

Given a pair of vectors under SU(2) $(\mathbf{A},\mathbf{B})$, then the representation $F_\pm$ acts on such a pair in the following way:

$$F_\pm(\mathbf{v},R):(\mathbf{A},\mathbf{B}) \to (R\mathbf{A}, R\mathbf{B} \pm \mathbf{v} \wedge R\mathbf{A}), \tag{2.7}$$

or, in general, F_λ, $\lambda \in \mathbf{R}^*$, can be defined by

$$F_\lambda(\mathbf{v},R):(\mathbf{A},\mathbf{B}) \to (R\mathbf{A}, R\mathbf{B} + \lambda\mathbf{v} \wedge R\mathbf{A}). \tag{2.8}$$

It is clear that the representations F_λ are equivalent for all the values of λ.

All the representations of G_0 so far defined, Δ_λ, D_λ, and F_λ are reducible but not completely reducible. The contragredient representation of Δ_λ is $D_{-\lambda}$. Thus if $(b,\mathbf{a})$ transforms under $\Delta_\pm$ and (β,α) under $D_\mp$, then $\beta b + \alpha\cdot\mathbf{a}$ is a scalar. Here, F_λ and its contragredient representation are equivalent. In fact, if $(\mathbf{A},\mathbf{B})$ supports the representation F_λ, then $(\mathbf{B},\mathbf{A})$ transforms under its contragredient representation. Therefore, when objects such as $(\mathbf{A},\mathbf{B})$ and $(\mathbf{A}',\mathbf{B}')$ are transformed under $F_\pm$, a scalar can be obtained in the following two ways:

$$\text{(i)}\ \mathbf{A}\cdot\mathbf{B}' + \mathbf{B}\cdot\mathbf{A}',$$
$$\text{(ii)}\ \mathbf{A}\cdot\mathbf{A}'. \tag{2.9}$$

For the objects $(b,\mathbf{a})$, $(b',\mathbf{a}')$, supporting the representations $\Delta_\pm$, we have that $((b,\mathbf{a})\odot(b,\mathbf{a}'),\mathbf{a}\odot\mathbf{a}')$ transforms under $F_\mp$, where $\mathbf{a}\odot\mathbf{a}' = \frac{1}{2}(\mathbf{a}\wedge\mathbf{a}')$, and $(b,\mathbf{a})\odot(b',\mathbf{a}') = \frac{1}{2}(\mathbf{a}b' - \mathbf{a}'b)$. Then, with a $\Delta_\pm$ object $(b,\mathbf{a})$, and an $F_\mp$ one $(\mathbf{A},\mathbf{B})$ we can build two kinds of scalars:

$$\text{(i)}\ (\mathbf{a}\odot\mathbf{a}')\cdot\mathbf{A} + ((b,\mathbf{a})\odot(b',\mathbf{a}'))\cdot\mathbf{B},$$
$$\text{(ii)}\ ((b,\mathbf{a})\odot(b',\mathbf{a}'))\cdot A. \tag{2.10}$$

Following the same arguments, other scalars can be obtained by means of $D_\pm$ objects.

B. The Galilean electromagnetism

Some years ago, Le Bellac and Lévy-Leblond[13] considered a nonrelativistic electromagnetic theory in agreement with the Galilean relativity. They found two possible nonrelativistic limits of the Maxwell equations: the magnetic limit (ml) and the electric limit (el). This Galilean electromagnetism make use of the representations of G_0 that we have just studied: $\Delta_\pm$, $D_\pm$, and $F_\pm$. A short description of both limits is given below.

1. The magnetic limit

The electromagnetic potential (emp), $(\varphi,\mathbf{A})$, is a four-vector under the representation $D_\pm$. Thus

$$(\mathbf{v},R):(\varphi,\mathbf{A})\to(\varphi\pm\mathbf{v}\cdot R\mathbf{A},R\mathbf{A}). \tag{2.11}$$

The electromagnetic field (emf) $f\equiv(\mathbf{B},\mathbf{E})$, related to this potential is

$$\mathbf{B}=\nabla\wedge\mathbf{A},\quad \mathbf{E}=-\nabla\varphi\mp\partial_t\mathbf{A}, \tag{2.12}$$

whence the emf transforms under $F_{\mp}$, as it is easy to check.

2. The electric limit

In this case, the emp is a $\Delta_{\pm}$ type four-vector:

$$(\mathbf{v},R):(\varphi,\mathbf{A})\to(\varphi,R\mathbf{A}\pm\mathbf{v}\varphi). \tag{2.13}$$

The associated emf $(\mathbf{E},\mathbf{B})$, is

$$\mathbf{E}=-\nabla\varphi,\quad \mathbf{B}=\nabla\wedge A, \tag{2.14}$$

which is an object which transforms under $F_{\pm}$.

As it is well known, the physical objects are the emf's and not the emp's. Therefore, it is not necessary that the emp transforms itself under the Galilei group according to the laws (2.11) or (2.13), to be sure that the emf does it in the correct way. It is sufficient that $\Delta_{\pm 1}(g)(\varphi,\mathbf{A})$ and $g(\varphi,\mathbf{A})$ differ in $(\varphi',\mathbf{A}')$ such that

$$-\nabla\varphi'\mp\partial_t\mathbf{A}'=\mathbf{0},\quad \nabla\wedge\mathbf{A}'=\mathbf{0}, \tag{2.15}$$

in the magnetic limit, or

$$-\nabla\varphi'=\mathbf{0},\quad \nabla\wedge\mathbf{A}'=\mathbf{0}, \tag{2.16}$$

in the electric case. This kind of transformation will be called "up to a gauge transformation."

On the other hand, if we want to include the parity operator $P:(t,\mathbf{x})\to(t,-\mathbf{x})$, as a symmetry, we need to fix how it has to be represented. In fact, there are two possible ways of enlarging the representations Δ, D, and F with P.

(a) In agreement with the classical theory, $\mathbf{E}$ has a vectorial character, while $\mathbf{B}$ a pseudovectorial one. Thus we can write

$$P:(\mathbf{E},\mathbf{B})\to(-\mathbf{E},\mathbf{B}), \tag{2.17}$$

while, to be consistent, the transformation of the potential is

$$P:(\varphi,\mathbf{A})\to(\varphi,-\mathbf{A}). \tag{2.17'}$$

(b) The other choice is given by

$$P:(\mathbf{E},\mathbf{B})\to(\mathbf{E},-\mathbf{B}),$$

$$P:(\varphi,\mathbf{A})\to(-\varphi,\mathbf{A}). \tag{2.18}$$

In this paper we will deal with constant emf's, and in this case for each limit we can choose the following potentials.

(1) Magnetic limit: Let $(\mathbf{B},\mathbf{E})$ be a constant emf, we can consider up to a gauge function the four potential $(\varphi,\mathbf{A})$ given by

$$\varphi=-\tfrac{1}{2}\mathbf{E}\cdot\mathbf{x},\quad \mathbf{A}=\tfrac{1}{2}(\mathbf{B}\wedge\mathbf{x}-\mathbf{E}t). \tag{2.19}$$

If $(\mathbf{B},\mathbf{E})$ transforms under F_-, then $(\varphi,\mathbf{A})$, so defined, does it under D_+, as one can check straightforwardly.

(2) Electric limit: Let $(\mathbf{E},\mathbf{B})$ be a constant emf. Here, the four-potential can be defined, up to a gauge, by

$$\varphi=-\mathbf{E}\cdot\mathbf{x},\quad \mathbf{A}=\tfrac{1}{2}\mathbf{B}\wedge\mathbf{x}. \tag{2.20}$$

Thus, if $(\mathbf{E},\mathbf{B})$ is an $F_{\pm}$ object, the potential $(\varphi,\mathbf{A})$ will be a $\Delta_{\pm}$ object, up to a gauge transformation $(\varphi',\mathbf{A}')$, which can never be reduced to zero. This is an important difference between the ml and the el whenever the fields are constant.

III. THE NONRELATIVISTIC MAXWELL GROUP

As in our paper I for the Poincaré group, the wave function describing an elementary system interacting with a constant emf, f_0 will be denoted by $\psi(x,f_0)$, where $x\in X$ (space-time manifold). This wave function belongs to a Hilbert space H_{f_0}. Let us take another observer and let g be the Galilean transformation relating both reference frames, then for this new observer the system will be described by $\psi'(x,f)$, $\forall x\in X$ with $f=gf_0$ where ψ' belongs to the Hilbert space H_f. Both descriptions must be physically equivalent, so we can write

$$\psi'(g(x,f))\equiv(U(g)\psi)(g(x,f))=A(g;x,f)\psi(x,f), \tag{3.1}$$

where (x,f) belongs to the orbit $\theta_{(x_0,f_0)}$ of the origin point (x_0,f_0), and A is a Borel nonsingular matrix-valued function, called gauge matrix, $A:G\times\theta_{(x_0,f_0)}\to\mathrm{GL}(n,\mathbf{C})$. The set of transformations $\{U(g)\}_{g\in G}$ is a generalization of the usual locally operating realizations.[1,3] In this case the transformations are functions of the constant emf as well as the space-time point.

Some properties of these realizations come from the fact that the product of two of them must give another transformation up to a phase ω depending in this case on the emf as well as on the group elements as usual, i.e.,

$$(U(g')U(g)\psi)(g'g(x,f))$$

$$=\omega(g',g;f)(U(g'g)\psi)(g'g(x,f)). \tag{3.2}$$

This relationship leads to the following one for the associated gauge matrices

$$A(g';g(x,f))A(g;x,f)=\omega(g',g;f)A(g'g;x,f). \tag{3.3}$$

The phase ω is called a factor system of G and it is a Borel function $\omega:G\times G\times\theta_f\to\mathrm{U}(1)$, θ_f is the orbit of the emf f under the action of G.

The associativity of these transformations allows us to prove that ω is a two-cocycle, i.e.,

$$\omega(g_3,g_2g_1;f)\omega(g_2,g_1;f)$$

$$=\omega(g_3,g_2;g_1f)\omega(g_3g_2,g_1;f). \tag{3.4}$$

Now we can define a relation of equivalence between local realizations. Two local realizations U and U' are locally pseudoequivalent if there are: (i) a local operator S defined in every $H_f\forall f\in\theta_{f_0}$, i.e.,

$$(S\psi)(x,f)=S(x,f)\psi(x,f),\quad S:\theta_{(x,f)}\to\mathrm{GL}(n,\mathbf{C})$$

being a Borel matrix-valued function and (ii) a Borel function $\lambda:G\times\theta_f\to U(1)$ verifying

$$(U'(g)\psi)(g(x,f))=\lambda(g,f)(SU(g)S^{-1}\psi)(g(x,f)). \tag{3.5}$$

The gauge matrices associated to local pseudoequivalent realizations are related by

$$A'(g;x,f)=\lambda(g,f)S(g(x,f))A(g;x,f)S^{-1}(x,f) \tag{3.6}$$

and the corresponding factor systems verify

$$\omega'(g',g;f)=\omega(g',g;f)\lambda(g',gf)\lambda(g,f)\lambda^{-1}(g'g,f). \tag{3.7}$$

The relationship (3.7) defines a relation of equivalence among the factor systems of G.

A. Factor systems of the Galilei group depending on a constant emf

The Galilei group acts on the manifolds F and $X\times F$, where F is the set of the constant emf's and X the space-time manifold. This action gives rise to orbits in both manifolds denoted by θ_{f_0} and $\theta_{(x_0,f_0)}$, respectively, where f_0 stands for $(\mathbf{E}_0,\mathbf{B}_0)$ or $(\mathbf{B}_0,\mathbf{E}_0)$ according to the limit of the emf. The isotopy groups of these orbits are Γ_{f_0} for $f_0\in\theta_{f_0}$ and $\Gamma_{(x_0,f_0)}$ for $(x_0,f_0)\in\theta_{(x_0,f_0)}$. Borel sections $s:\theta_{(x_0,f_0)}\to G$ and $r:\theta_{f_0}\to G_0$ can be defined by

$$s(x,f)=(x,\Lambda_f),\ \text{with}\ (x,\Lambda_f):(x_0,f_0)\to(x,f) \tag{3.8}$$

and

$$r(f)=\Lambda_f, \tag{3.9}$$

where $x\in T_4$, $\Lambda_f\in G_0$. Thus, any $g\in G$ can be decomposed as

$$g=(a,\Lambda)=(a,\Lambda_{gf_0})(0,\gamma(g)), \tag{3.10}$$

if we make use of the first section, and

$$g=(a,\Lambda)=(0,\Lambda_{gf_0})(\Lambda_{gf_0}^{-1}a,\gamma(g)),\ \text{where}\ \gamma(g)\in\Gamma_{(x_0,f_0)} \tag{3.10'}$$

when we try the second one.

In order to compute the factor systems of the Galilei group depending on a constant emf, we will follow the same method used in I, so we will not give a thorough exposition and we will restrict ourselves to the main points. These are given in three steps.

1. The orbits of the Galilean emf's and their little groups

a. Magnetic limit. There are three nonzero kinds of orbits:[14]

(I) $(\mathbf{B},\mathbf{E})$ orbits. There exits a unique element $(\hat{\mathbf{B}},\hat{\mathbf{E}})$ in each of these orbits such that $\hat{\mathbf{B}}=(0,0,\hat{B})$, $\hat{\mathbf{E}}=(0,0,\hat{E})$, $\hat{B}\in\mathbf{R}^+$, $\hat{E}\in\mathbf{R}^*$. They are four dimensional and their characterizing invariants are $\mathbf{B}^2=C_1>0$ and $\mathbf{B}\cdot\mathbf{E}=C_2\neq 0$. The little group of a point $(\mathbf{B},\mathbf{E})$ is generated by the space-time translations together with $\mathbf{J}\cdot\mathbf{B}-\mathbf{K}\cdot\mathbf{E}$ and $\mathbf{K}\cdot\mathbf{B}$. For the particular emf $(\hat{\mathbf{B}},\hat{\mathbf{E}})$ the two last generators are J_z and K_z.

(II) $(\mathbf{B},\mathbf{0})$ orbits. A characteristic element is $\hat{\mathbf{B}}=(0,0,\hat{B})$ and $\hat{\mathbf{E}}=0$, $\hat{B}\in\mathbf{R}^+$. These orbits belong to the same strata of the orbits of the type I, but here $\mathbf{B}\cdot\mathbf{E}=C_2=0$.

(III) $(\mathbf{0},\mathbf{E})$ orbits. We can choose $\mathbf{B}=\mathbf{0}$, $\hat{\mathbf{E}}=(0,0,\hat{E})$, $\hat{E}\in\mathbf{R}^+$. The orbits are two dimensional, and their invariant is $\mathbf{E}^2=C_3>0$. The isotopy group of $(\mathbf{0},\mathbf{E})$ is generated by the four space-time translation generators, $\mathbf{J}\cdot\hat{\mathbf{E}}$ and $\mathbf{K}$. If we take the point $(\mathbf{0},\hat{\mathbf{E}})$ then the generator $\mathbf{J}\cdot\hat{\mathbf{E}}$ is proportional to J_z.

b. Electric limit.[4] The results are similar to the magnetic case, if we interchange the roles of $\mathbf{B}$ and $\mathbf{E}$.

2. Extensions of the groups Γ_f

a. Magnetic limit.[4] Type I and II orbits $[f=(\mathbf{B},\mathbf{E})]$. The group Γ_f is $T_4\odot\Gamma_{(x_0,f)}$, where T_4 is the group of the space-time translations, and $\Gamma_{(x_0,f)}$ is generated by J_z and K_z. The second cohomology group $H^2(\Gamma_f,\mathrm{U}(1))$ is isomorphic to $\mathbf{R}^3$, which means that the space of equivalence classes of central extensions of Γ_f by $\mathrm{U}(1)$ is three dimensional. A representative of each class of these central extensions has the new nonvanishing commutators of its Lie algebra given by

$$[H,P_z]=-\epsilon,\quad [P_x,P_y]=\beta,$$

$$[P_z,K_z]=-m,\quad \epsilon,\beta,m\in\mathbf{R}^*$$

and a lifting of an element $[\epsilon,\beta,m]\in H^2(\Gamma_f,\mathrm{U}(1))$ is the following factor system belonging to $Z^2(\Gamma_f,\mathrm{U}(1))$:

$$\omega_{\epsilon,\beta,m}(g',g)=\omega_\epsilon(g',g)\omega_\beta(g',g)\omega_m(g',g),\quad g',g\in\Gamma_f,$$

where

$$\omega_\epsilon(g',g)=\exp\{i\epsilon(b',a'_z)\odot(v_z)(b,a_z)\},$$

$$\omega_\beta(g',g)=\exp\{i\beta(\mathbf{a}'_\perp\odot R_z\mathbf{a}_\perp)\},$$

$$\omega_m(g',g)=\exp\{im(\tfrac{1}{2}v'^2_z b+v'_z(R\mathbf{a})_z)\},$$

where g', $g\in\Gamma_f$, $g\equiv(b,\mathbf{a},v_z,R_z)$, $\mathbf{a}=(\mathbf{a}_\perp,a_z)\equiv(a_1,a_2,a_3)$, with $(v_z)(b,a_z)=(b,a_z+v_z b)$.

Type III orbits $[F=(\mathbf{0},\mathbf{E})]$. Here, $\Gamma_f=T_4\odot\Gamma_{(x_0,f)}\equiv T_4\odot\langle J_z,\mathbf{K}\rangle$ and $H^2(\Gamma_f\mathrm{U}(1))\simeq\mathrm{R}^4$. A lifting of an element $[m_\parallel,m_\perp,\epsilon,\lambda]\in H^2(\Gamma_f,\mathrm{U}(2))$ is the following one:

$$\omega_{m\parallel,m\perp,\epsilon,\lambda}(g',g)$$

$$=\omega_{m\parallel}(g',g)\omega_{m\perp}(g',g)\omega_\epsilon(g',g)\omega_\lambda(g',g), \tag{3.11}$$

where

$$\omega_{m\parallel}(g',g)=\exp\{im_\parallel(\tfrac{1}{2}v'^2_z b+v'_z a_z)\},$$

$$\omega_{m\perp}(g',g)=\exp\{im_\perp(\tfrac{1}{2}\mathbf{v}'^2_\perp b+\mathbf{v}'_\perp\cdot R_z\mathbf{a}_\perp)\},$$

$$\omega_\epsilon(g',g)=\exp\{i\epsilon[(b',a'_z)\odot(v_z)(b,a_z)]\}, \tag{3.12}$$

$$\omega_\lambda(g',g)=\exp\{i\lambda[\mathbf{v}'_\perp\odot R_z\mathbf{v}_\perp]\},$$

where $g=(b,\mathbf{a},\mathbf{v},R_z)\in\Gamma_f$, $\mathbf{a}=(\mathbf{a}_\perp,a_z)$, $\mathbf{v}=(\mathbf{v}_\perp,v_z)$.

According to the theory of locally operating realizations the factor system ω_λ cannot appear in this kind of realizations. Consequently, we will consider $\lambda = 0$, i.e., $\omega_\lambda = 1$ in the sequel. The other parameters $m_\parallel$, $m_\perp$, and ϵ are related with the following nonzero Lie commutators:

$$[P_x,K_x] = [P_y,K_y] = -m_\perp, \quad [P_z,K_z] = -m_\parallel,$$
$$[H,P_z] = -\epsilon. \tag{3.13}$$

b. Electric limit. The results are the same of the earlier case for the three types of orbits provided we interchange **E** and **B**.

3. Factor systems of the whole Galilei group

Now, we compute the factor systems $\omega(f;g,g')$ of G using the factor systems of the isotopy groups Γ_f and the properties which are given in I. Here, we quote two of such properties which are useful enough for our purposes (where we have changed slightly the notation with respect to I in order to simplify the final results):

$$\text{(i)}\ \omega(f_0;gr(f),g'r(f')) \tag{3.14}$$
$$= \omega(f_0;g,\delta(r(f)g'r(f'))),$$

with g, $g' \in \Gamma_{f_0}$, $\delta \in \Gamma_{f_0}$, and the normalized section $r: G/\Gamma_{f_0} \to G_0 \subset G$,

$$\text{(ii)}\ \omega(f;g,g') = \omega(f_0;\Lambda^{-1}{}_f g,g'),$$

where g, $g' \in G$, $f_0 \equiv \hat{f}$, $\Lambda_f \in G_0$ such that $\Lambda_f f_0 = f$ and $r(f) = \Lambda_f^{-1}$.

We distinguish between the two limits.

a. Magnetic limit. Type I and II orbits. Here, we obtain

$$\omega(\Lambda f;x,x') = \omega(f;\Lambda^{-1}x,\Lambda^{-1}x'), \quad \forall f \in F,\ \forall x,x' \in X,\ \forall \Lambda \in G_0 \tag{3.15}$$

i.e., ω is a scalar in f and (x,x'). However, according to Sec. II, the possible independent scalars made out of f and (x,x') are

$$\text{(1)}\ (\mathbf{x} \odot \mathbf{x}')\cdot\mathbf{B} + ((t,\mathbf{x}) \odot (t',\mathbf{x}'))\cdot\mathbf{E}, \tag{3.16}$$
$$\text{(2)}\ ((t,\mathbf{x}) \odot (t',\mathbf{x}'))\cdot\mathbf{B},$$

where $(t,\mathbf{x}) \odot (t',\mathbf{x}') = \frac{1}{2}(\mathbf{x}t' - \mathbf{x}'t)$. Thus a function of these scalars restricted to Γ_f must agree up to equivalence with the exponents of the factor systems ω_ϵ and ω_β. This implies that this function must be linear in both scalars (1) and (2). Moreover, taking into account the parity, the only admissible scalar is given by (1) and we obtain that

$$\omega(f_0;x,x') \equiv \omega_\epsilon(x,x')\omega_\beta(x,x') = \exp\{iq[(\mathbf{x} \odot \mathbf{x}')\cdot\hat{\mathbf{B}} + ((t,\mathbf{x}) \odot (t',\mathbf{x}'))\cdot\hat{\mathbf{E}}]\}, \tag{3.17}$$

whence, identifying the exponents of this equality we get that $\epsilon = q\hat{E}$ and $\beta = q\hat{B}$. The real number q, which is a characteristic of the interacting system, will be called the electric charge.

Finally, the generic factor system originated by ω_ϵ and ω_β of Γ_f is

$$\omega_q(f;g,g') = \exp\{iq[(\mathbf{x} \odot (\mathbf{v},R)_s(t',\mathbf{x}'))\cdot\mathbf{B} + ((t,\mathbf{x}) \odot (\mathbf{v},R)(t',\mathbf{x}'))\cdot\mathbf{E}]\}, \tag{3.18}$$

with the implicit notation $(\mathbf{v},R)_s(t',\mathbf{x}') = R\mathbf{x}' + \mathbf{v}t'$. The other factor system ω_m corresponds to the restriction to Γ_f of the well-known extension of the Galilei $(3+1)$ group given in Sec. I, and it is related with the mass of the system. However, another factor system $\omega_{m,m'}$ of G, equivalent to ω_m can be written as

$$\omega_{m,m'}(f;g,g') = \exp\{im(\tfrac{1}{2}\mathbf{v}^2 b' + \mathbf{v}\cdot R\mathbf{a}') + im'[\tfrac{1}{2}(\mathbf{v}\cdot\mathbf{B})^2 b' + (\mathbf{v}\cdot\mathbf{B})(\mathbf{B}\cdot R\mathbf{a}')]\}. \tag{3.19}$$

The meaning of m' will be discussed later, in connection with the orbits of type III.

Orbits III. The factor systems linked with the emf take the form (3.18) if we make $\mathbf{B} = \mathbf{0}$ ($\beta = 0$), while the factor systems related with the masses $m_\parallel$ and $m_\perp$ can be written in terms of m and m' as

$$\omega_m(g,g') = \exp\{im(\tfrac{1}{2}\mathbf{v}^2 b' + \mathbf{v}\cdot R\mathbf{a}')\},$$
$$\omega_{m'}(f;g,g') = \exp\{im'[\tfrac{1}{2}(\mathbf{v}\cdot\mathbf{E})^2 b' + (\mathbf{v}\cdot\mathbf{E})(\mathbf{E}\cdot R\mathbf{a}')]\}. \tag{3.20}$$

The former corresponds to the known extension of G by U(1) (independent of the emf), and m' can be associated to an "anisotropic" mass of the system.

b. Electric limit. For the orbits I′ and II′ we have two possible scalars as in (3.16) except that it must be interchanged the roles of **B** and **E**. Following the same justification of the magnetic limit we obtain

$$\omega(f;x,x') = \exp\{iq((t,\mathbf{x}) \odot (t',\mathbf{x}'))\cdot E\} \tag{3.21}$$

and identifying this exponent with the two exponents of Γ_f when (3.21) is restricted to this subgroup, we obtain $\beta = 0$ and $\epsilon = qE$. Thus the general expression of these factor systems are

$$\omega'_q(f;g,g') = \exp\{iq[(t,\mathbf{x}) \odot (\mathbf{v},R)(t',\mathbf{x}')]\cdot\mathbf{E}\}, \tag{3.22}$$

in which it has lost any track of the magnetic field. The other factor system ω_m is the corresponding one to the central extension of G by U(1).

Orbits III′. We obtain similar results to the corresponding III orbits if we replace **B** by **E** and β by ϵ, except that here the only surviving factor systems are $\omega'_{m\parallel}$ and $\omega_{m\perp}$.

The appearance of an anisotropic mass in orbits III and III′ can be explained by the existence of a privileged direction defined by **E** or **B**, respectively. This kind of orbit has few signs of "electromagnetic character." There exists only an electric field **E** (III) or a magnetic field **B** (III′), which transform as vectors under rotations. This is not a

realistic situation if the true kinematical group is the Poincaré group. The Galilean electromagnetism was obtained by taking the magnetic limit (or electric limit) with $|\mathbf{E}| \ll q|\mathbf{B}|$ ($|\mathbf{B}| \ll q|\mathbf{E}|$). However, if $|\mathbf{B}| = 0$ ($|\mathbf{E}| = 0$) the limit is not evidently correct. If we consider the electric field of orbit III as the limit of $f(\mathbf{E},\mathbf{B})$ when $\mathbf{B} \to \mathbf{0}$, and take the same limit for the factor systems we do not get the factor system of the anisotropic mass. A similar reasoning applies to orbits III'. For all these reasons hereafter, we will not consider the anisotropic mass factor systems, and also the parameter q will be assumed to be independent of the emf's orbits.

Note that the factor system $\omega_{m'}$ also appears in orbits I and II (I' and II'), but this time the isotopy group is smaller and as a consequence, the equivalence being wider, $\omega_{m',m}$ is equivalent to ω_m.

B. A representation group for the Galilei group: The nonrelativistic Maxwell group

According to the two kinds of limits we can consider two possible Maxwell groups, one for each limit.

1. The magnetic limit

In the following we will choose $q = 1$. Let $V(\theta_f)$ be the set of the real linear functions defined on the orbit θ_f by

$$[\mathbf{j},\mathbf{k}](\mathbf{B},\mathbf{E}) = \mathbf{j}\cdot\mathbf{B} + \mathbf{k}\cdot\mathbf{E}, \quad \text{with } [\mathbf{j},\mathbf{k}] \in V(\theta_f). \tag{3.23}$$

This function space has the structure of an Abelian Lie group ($\simeq \mathbf{R}^6$). We define the nonrelativistic Maxwell group for the magnetic limit M_m as an extension of the extended Galilei group ($m = 1$) by $V(\theta_f)$ with the following composition law:

$$([\mathbf{j}',\mathbf{k}'],\theta',b',\mathbf{a}',\mathbf{v}',R')([\mathbf{j},\mathbf{k}],\theta,b,\mathbf{a},\mathbf{v},R)$$
$$= ([\mathbf{j}',\mathbf{k}'] + (\mathbf{v}',R')[\mathbf{j},\mathbf{k}] + W_1(g',g),\theta' + \theta + W_2(g',g),b' + b,\mathbf{a}' + R'\mathbf{a} + \mathbf{v}'b,\mathbf{v}' + R'\mathbf{v},R'R), \tag{3.24}$$

where $g = (b,\mathbf{a},\mathbf{v},R)$, $((\mathbf{v},R)[\mathbf{j},\mathbf{k}])(\mathbf{B},\mathbf{E}) = [\mathbf{j},\mathbf{k}]((\mathbf{v},R)^{-1}(\mathbf{B},\mathbf{E}))$,

$$W_1(g',g) = [-\mathbf{a}'\odot(\mathbf{v}',R')_s(b,\mathbf{a}), -(b',\mathbf{a}')\odot(\mathbf{v}',R')(b,\mathbf{a})],$$

and

$$W_2(g',g) = \tfrac{1}{2}\mathbf{v}'^2 b + \mathbf{v}'\cdot R'\mathbf{a}.$$

The Lie algebra is 17 dimensional and besides the Lie commutators of the Galilei group we have the following nonzero commutators:

$$[K_i,P_j] = \delta_{ij}I, \quad [P_i,P_j] = -\epsilon_{ijk}\beta_k, \quad [H,P_i] = \epsilon_i,$$
$$[J_i,\beta_j] = \epsilon_{ijk}\beta_k, \quad [J_i,\epsilon_j] = \epsilon_{ijk}\epsilon_k, \quad [K_i,\epsilon_j] = -\epsilon_{ijk}\beta_k, \tag{3.25}$$

where I, β, and ϵ are the generators associated to the subgroups with parameters θ, $\mathbf{j}$, and $\mathbf{k}$, respectively.

2. The electric limit

Following the same procedure of the earlier case, the nonrelativistic Maxwell group in the electric limit, M_e, is an extension of the extended Galilei group by the Abelian group $V(\theta_f)$ of the functions $\mathbf{k}$, defined by $\mathbf{k}(\mathbf{B},\mathbf{E}) = \mathbf{k}\cdot\mathbf{E}$, with the composition law:

$$(\mathbf{k}',\theta',g')(\mathbf{k},\theta,g) = (\mathbf{k}' + (\mathbf{v}',R')\mathbf{k} + W_1(g',g),\theta' + \theta + W_2(g',g),g'g) \tag{3.26}$$

with g', $g \in G$, $(\mathbf{v},R)\mathbf{k} = R\mathbf{k}$,

$$W_1(g',g) = -(b',\mathbf{a}')\odot(\mathbf{v}',R')(b,\mathbf{a})$$

and

$$W_2(g',g) = \tfrac{1}{2}\mathbf{v}'^2 b + \mathbf{v}'\cdot R'\mathbf{a}.$$

Its Lie algebra has 14 generators. The following Lie commutators together with the commutators of the Galilei group given in (2.1) constitute its nonzero commutators:

$$[K_i,P_j] = \delta_{ij}I, \quad [H,P_i] = \epsilon_i, \quad [J_i,\epsilon_j] = \epsilon_{ijk}\epsilon_k, \tag{3.27}$$

where I and ϵ are the generators associated to the parameters θ and $\mathbf{k}$, respectively.

IV. THE IRREDUCIBLE REPRESENTATIONS OF THE MAXWELL GROUP

In this section we will build up the irreducible representations of the two Maxwell groups M_m and M_e following the Kirillov–Mackey method. In I, the interested reader can find a short explanation of this method and see how it works in the case of the relativistic Maxwell group.

A. The magnetic limit: M_m

1. The computation of the orbits

The group M_m can be decomposed as a semidirect product $M_m = N \odot H$, where $N = \{([\mathbf{j},\mathbf{k}],\theta,b,\mathbf{a})\}$ is a nilpotent group and $H = \{(\mathbf{v},R)\}$. Let $\mathcal{N}$ be the Lie algebra associated to the group N and $\mathcal{N}^*$ the dual space of $\mathcal{N}$. The coadjoint action of N on $\mathcal{N}^*$ is defined by

$$\langle \mathrm{coad}_n A^*, A\rangle = \langle A^*, \mathrm{ad}_n^{-1} A\rangle, \quad \text{with } A \in \mathcal{N}, A^* \in \mathcal{N}^*, \text{ and } n \in N.$$

Let $([\mathbf{B},\mathbf{E}],m,\epsilon,\mathbf{p})$ be the coordinates of a generic element of $\mathcal{N}^*$ and $n \equiv ([j,\mathbf{k}],\theta,b,\mathbf{a}) \in N$. Then the coadjoint action of n on $\mathcal{N}^*$ is given by ($A'^* = \mathrm{coad}_n A^*$):

$$[\mathbf{B}',\mathbf{E}'] = [\mathbf{B},\mathbf{E}], \quad m' = m,$$
$$\epsilon' = \epsilon + \mathbf{E}\cdot\mathbf{a}, \quad \mathbf{p}' = \mathbf{p} - \mathbf{E}b + \mathbf{B}\wedge\mathbf{a}. \tag{4.1}$$

This action splits $\mathcal{N}^*$ into orbits that are classified according to the types of the emf's given in Sec. III A.

(i) Parallel type (**B**,**E**). In this case the orbits θ are four dimensional and are characterized by (**B**,**E**) and m. One of this kind of orbits will be denoted by $\theta([\mathbf{B},\mathbf{E}],\mathrm{m})$.

(ii) Magnetic type (**B**). These orbits, $\theta([\mathbf{B},\mathbf{E}],m,\epsilon,p_\parallel)$ have dimension two. The invariants are (**B**,**E**), m, ϵ, and $\mathbf{p}\cdot\mathbf{B} = p_\parallel$.

(iii) Electric type (**E**). The invariants of these orbits are (**0**,**E**), m, and $\mathbf{p}\wedge\mathbf{E}=\mathbf{p}_\perp$. The orbits $\theta([\mathbf{0},\mathbf{E}],m,\mathbf{p}_\perp)$ are two dimensional.

(iv) Null type (**B**,**E**) = (**0**,**0**). Each of these orbits has only one point $((\mathbf{0},\mathbf{0}),m,\epsilon,\mathbf{p})$.

Now, if we fix a point, u, in each orbit θ and compute a maximal subalgebra $\mathscr{K}\subset\mathscr{N}$ such that $[\mathscr{K},\mathscr{K}]\subset\mathrm{Ker}(u)$, then a unitary representation U of the Lie group K is given by

$$U_u:k\to\exp\{i\chi_u(k)\},\tag{4.2}$$

where $\chi_u(k)=\langle u,k\rangle$, being $k\in\mathscr{K}$ the corresponding Lie algebra element associated to a group element of K. These representations induce unitary irreducible representations of N. If we take two points of a same orbit the induced representations linked to these two points are equivalent. Thus the equivalence classes of these unitary irreducible representations (uir) are in one-to-one correspondence with the orbits of $\mathscr{N}^*$ under the action of N.

The group H acts on the set of classes of equivalence of the uir's of N, i.e., on the set of orbits θ of $\mathscr{N}^*$ under N. The action of H splits the space of the orbits θ into orbits that we will call superorbits. If we choose a representative of each orbit, this action is given by

$$\begin{aligned}&(\mathbf{v},R):([\mathbf{B},\mathbf{E}],m,\epsilon,\mathbf{p})\\&\quad\to([R\mathbf{B},R\mathbf{E}-\mathbf{v}\wedge R\mathbf{B}],m,\epsilon+\tfrac{1}{2}m\mathbf{v}^2\\&\quad md;dis;-\mathbf{v}\cdot R\mathbf{p},R\mathbf{p}-m\mathbf{v}).\end{aligned}\tag{4.3}$$

The classification of these superorbits is the following.

(I) Superorbits of type (**B**,**E**). The invariants are m, $\mathbf{B}^2=C_1>0$, $\mathbf{B}\cdot\mathbf{E}=C_2\neq0$. We denote these superorbits by

$$\Theta([\mathbf{B},\mathbf{E}],m)=\bigcup_{\substack{\mathbf{B}^2=C_1>0\\ \mathbf{B}\cdot\mathbf{E}=C_2\neq0}}\theta([\mathbf{B},\mathbf{E}],m).$$

The isotopy group of each of these orbits is two dimensional. In particular, the one of $\theta([\hat{\mathbf{B}},\hat{\mathbf{E}}],m)$ is generated by J_z and K_z ($\hat{\mathbf{B}}=(0,0,\hat{B})$, $\hat{\mathbf{E}}=(0,0,\hat{E})$).

(II) Superorbits of type (**B**). Now the invariants are m, $\mathbf{B}^2=C_1>0$, $\mathbf{B}\cdot\mathbf{E}=0$, and $(\mathbf{p}_\parallel{}^2/2m)-\epsilon=\lambda$, $\lambda\in\mathbf{R}$. Thus these superorbits are five dimensional:

$$\Theta([\mathbf{B},\mathbf{E}],m,\lambda)=\bigcup_{\substack{\mathbf{B}^2=C_1>0\\ \mathbf{B}\cdot\mathbf{E}=0\\ (\mathbf{p}_\parallel{}^2/2m)-\epsilon=\lambda}}\theta([\mathbf{B},\mathbf{E}],m,\epsilon,\mathbf{p}_\parallel).$$

The isotopy group is one dimensional. In particular, for the element $\theta([\hat{\mathbf{B}},\hat{\mathbf{E}}],m,\epsilon,p_\parallel)$ the isotopy group is generated by J_z.

(III) Superorbits of type (**E**). In this case the invariants are $\mathbf{E}^2$ and m. The dimension of these superorbits is four, whence the isotopy group of any of its orbits is two dimensional:

$$\Theta([\mathbf{0},\mathbf{E}],m)=\bigcup_{\mathbf{p}_\perp,\mathbf{E}^2=C_1'}\theta([\mathbf{0},\mathbf{E}],m,\mathbf{p}_\perp).$$

The isotopy group of the characteristic element is generated by J_z and K_z.

2. The computation of the little groups

The above classifications into orbits and superorbits will be useful in the computation of the "orbital invariants" that label the uir's of M_m. To complete the characterization of each uir of M_m we now compute the generators of the little group of an element belonging to each kind of superorbit. These generators will constitute the "little group invariants." Let $([\mathbf{B},\mathbf{E}],m,\epsilon,\mathbf{p})$ be a generic point of a superorbit, the action of an element $(\mathbf{v},R)$ of H on $[\mathbf{B},\mathbf{E}]$ is $(\mathbf{v}_1R,)[\mathbf{B},\mathbf{E}]=[R\mathbf{B},R\mathbf{E}-\mathbf{v}\wedge\mathbf{B}]$. The action of an infinitesimal transformation $(\mathbf{v},\alpha)$ is

$$(\mathbf{v},\alpha):[\mathbf{B},\mathbf{E}]\to[\mathbf{B}+\alpha\wedge\mathbf{B},\mathbf{E}+\alpha\wedge\mathbf{E}-\mathbf{v}\wedge\mathbf{B}].\tag{4.4}$$

The solutions that leave invariant $[\mathbf{B},\mathbf{E}]$ are

$$\begin{aligned}&(i)\alpha=0,\ \mathbf{v}\propto\mathbf{B},\text{ which correspond to the}\\&\quad\text{generator }\mathbf{B}\cdot\mathbf{K},\\&(ii)(\alpha,\mathbf{v})\propto(\mathbf{B},-\mathbf{E}),\text{ whose generator}\\&\quad\text{is }\mathbf{B}\cdot\mathbf{J}-\mathbf{E}\cdot\mathbf{K}.\end{aligned}\tag{4.5}$$

Now, we analyze each superorbit in more detail.

a. Superorbits of type (B,E). The infinitesimal action of the uniparametric subgroups generated by each of the generators (4.5) is

$$\mathbf{B}\cdot\mathbf{K}:([\mathbf{B},\mathbf{E}],m,\epsilon,\mathbf{p})\to([\mathbf{B},\mathbf{E}],m,\epsilon-\mathbf{B}\cdot\mathbf{p},\mathbf{p}-m\mathbf{B})\tag{4.6}$$

$$\begin{aligned}&\mathbf{B}\cdot\mathbf{J}-\mathbf{E}\cdot\mathbf{K}:([\mathbf{B},\mathbf{E}],m,\epsilon,\mathbf{p})\\&\quad\to([\mathbf{B},\mathbf{E}],m,\epsilon+\mathbf{E}\cdot\mathbf{p},\mathbf{p}+\mathbf{B}\wedge\mathbf{p}+m\mathbf{E}).\end{aligned}\tag{4.7}$$

The action of an element $n\in N$ will be

$$\begin{aligned}&([\mathbf{j},\mathbf{k}],\theta,b,\mathbf{a}):([\mathbf{B},\mathbf{E}],m,\epsilon,\mathbf{p})\\&\quad\to([\mathbf{B},\mathbf{E}],m,\epsilon+\mathbf{E}\cdot\mathbf{a},\mathbf{p}-\mathbf{E}b+\mathbf{B}\wedge\mathbf{a}).\end{aligned}\tag{4.8}$$

If we require that the point remains invariant under such group actions, it follows from (4.6) and (4.8) that

$$\mathbf{B}\cdot\mathbf{p}=\mathbf{E}\cdot\mathbf{a},\tag{4.9}$$

$$m\mathbf{B}=-\mathbf{E}b+\mathbf{B}\wedge\mathbf{a}.\tag{4.10}$$

Multiplying (4.10) by **B** we obtain $b=-(m\mathbf{B}^2)/(\mathbf{E}\cdot\mathbf{B})$. Therefore, (4.10) is rewritten as

$$\mathbf{B}\wedge\mathbf{a}=m\mathbf{B}-[m\mathbf{B}^2/(\mathbf{E}\cdot\mathbf{B})]\mathbf{E}.\tag{4.10'}$$

This means that $\mathbf{a}=\alpha\mathbf{B}+\beta\mathbf{E}\wedge\mathbf{B}$. Putting this value of **a** in (4.9), we obtain $\alpha=(\mathbf{B}\cdot\mathbf{p})/(\mathbf{E}\cdot\mathbf{B})$. The substitution of these values of **a** and α in (4.10) give after some calculations the result $\beta=-m/(\mathbf{E}\cdot\mathbf{B})$. Thus finally

$$\mathbf{a}=[(\mathbf{B}\cdot\mathbf{p})/(\mathbf{E}\cdot\mathbf{B})]\mathbf{B}-[m/(\mathbf{E}\cdot\mathbf{B})]\mathbf{E}\wedge\mathbf{B}.\tag{4.11}$$

The corresponding generator, $\mathbf{v}\cdot\mathbf{K}+\theta\cdot\mathbf{J}+bH+\mathbf{a}\cdot\mathbf{P}$, of the little group will be

$$\mathbf{B}\cdot\mathbf{K}-m\frac{\mathbf{B}^2}{\mathbf{E}\cdot\mathbf{B}}H+\frac{\mathbf{B}\cdot\mathbf{p}}{\mathbf{E}\cdot\mathbf{B}}\mathbf{B}\cdot\mathbf{P}-\frac{m}{\mathbf{E}\cdot\mathbf{B}}(\mathbf{E}\wedge\mathbf{B})\cdot\mathbf{P},\tag{4.12}$$

that can be rewritten as

$$H+\frac{i(\mathbf{B}\cdot\mathbf{P})^2}{2m\mathbf{B}^2}+\frac{(\mathbf{E},\mathbf{B},\mathbf{P})}{\mathbf{B}^2}-\frac{\mathbf{E}\cdot\mathbf{B}}{m\mathbf{B}^2}\mathbf{B}\cdot\mathbf{K}=i\lambda_1, \tag{4.12'}$$

where $i\lambda_1$ denotes its eigenvalue. Now requiring that (4.7) and (4.8) leave invariant one point and making a similar reasoning, we obtain the expression for the second generator

$$-\mathbf{B}\cdot\mathbf{J}+\mathbf{E}\cdot\mathbf{K}-mH+\mathbf{p}\cdot\mathbf{P}, \tag{4.13}$$

or

$$H+i\mathbf{P}^2/2m+(\mathbf{B}\cdot\mathbf{J}-\mathbf{E}\cdot\mathbf{K})/m=i\lambda_2, \tag{4.13'}$$

where $i\lambda_2$ is the eigenvalue. Both (4.12') and (4.13') are Casimirs that characterize the uir's of M. However, since a rotation of 2π must be represented by ± 1, the representation of the operators $\mathbf{B}\cdot\mathbf{K}$ and $\mathbf{B}\cdot\mathbf{J}-\mathbf{E}\cdot\mathbf{K}$ have to assure this condition and then λ_1 and λ_2 cannot be arbitrary real numbers. In fact it is easy to see that they are fixed by one real and one half-integer number.

b. Superorbits of type (B). In this case the isotopy group is one dimensional and its generator is computed from Eqs. (4.7) and (4.8) in a similar way to the above cases. We get

$$H+i\mathbf{P}^2/2m+(\mathbf{B}\cdot\mathbf{J}-\mathbf{E}\cdot\mathbf{K})/m=i\lambda_2. \tag{4.14}$$

The eigenvalue λ_2 can take half-integer values only. We must recall that one of the orbital invariants for this case is $\epsilon - p_{\parallel}^2/2m$, or in terms of generators: $H + i(\mathbf{P}\cdot\mathbf{B})^2/(2m\mathbf{B}^2) = i\lambda_1$, which has the same form as the invariant (4.12') of the previous superorbit.

c. Superorbits of type (E). Here, the isotopy group of any of its points is two dimensional. The generators of this isotopy group can be found in the same way as we did before for the superorbits of type I by replacing $\mathbf{B}\cdot\mathbf{K}$ by $\mathbf{E}\cdot\mathbf{J}$ and $\mathbf{B}\cdot\mathbf{J}-\mathbf{E}\cdot\mathbf{K}$ by $\mathbf{v}\cdot\mathbf{K}$ in (4.6), and the final result is

$$H+i\mathbf{P}^2/2m-(\mathbf{E}\cdot\mathbf{K})/m=i\lambda'_2,$$

$$\mathbf{E}\cdot\mathbf{J}-(\mathbf{E},\mathbf{P},\mathbf{K})/m=i\lambda'_1. \tag{4.15}$$

In this case the remark made for the superorbits of type I about the semi-integer character of the eigenvalues of the rotation generator is also valid, but here it affects to λ'_1 only.

With respect to these results we are going to make the following remarks: (i) In all the three superorbits, the invariant $\lambda_2(\lambda_2')$ has the same functional form. (ii) Nevertheless, the invariant λ_1 does not have the same form because of the superorbits of type III. Its character is also different: for superorbits I and III it is a generator of the little group of a point, while in the case II is an orbital invariant. (iii) The physical meaning of these invariants will be shown in the next section, when we construct the local representations of the Maxwell group.

B. The electric limit: M_e

This group M_e also has the structure of a semidirect product: $M_e = N\odot H$, where $N=\{(\mathbf{k},\theta,b,\mathbf{a})\}$ and $H=\{(\mathbf{v},R)\}$. Let $(\mathbf{E},m,\epsilon,\mathbf{p})$ be a generic element of $\mathcal{N}^*$. The coadjoint action of N over $\mathcal{N}^*$ is given by

$$(\mathbf{k},\theta,b,\mathbf{a}):(\mathbf{E},m,\epsilon,\mathbf{p})\to(\mathbf{E},m,\epsilon+\mathbf{E}\cdot\mathbf{a},\mathbf{p}-\mathbf{E}b). \tag{4.16}$$

The action of the subgroup H over $\mathcal{N}^*$ is the following one

$$(\mathbf{v},R):(\mathbf{E},m,\epsilon,\mathbf{p})\to(R\mathbf{E},m,\epsilon+\tfrac{1}{2}m\mathbf{v}^2-\mathbf{v}\cdot R\mathbf{p},R\mathbf{p}\cdot m\mathbf{v}). \tag{4.17}$$

The action of N on $\mathcal{N}^*$ splits it into the orbits:

(i) $\mathbf{E}=\mathbf{0}$, then each orbit is a point $(\mathbf{0},m,\epsilon,\mathbf{p})$.

(ii) $\mathbf{E}\neq\mathbf{0}$, the invariants of one of these orbits are $\mathbf{E},m,\mathbf{p}\wedge\mathbf{E}=\mathbf{p}_\perp$.

These orbits are two dimensional and are denoted by $\theta(\mathbf{E},\epsilon,\mathbf{p}_\perp)$.

The case (i) is not of interest because it reduces to a free system. As we said before in the magnetic case, the classes of equivalence of the induced representations of N are in a one-to-one correspondence with the orbits of $\mathcal{N}^*$ under N. The action of H in the set of orbits of $\mathcal{N}^*$ give rise to the superorbits. The invariants of this action for each superorbit are E^2 and m. The results for the little group invariants in this case are the same as in the magnetic superorbits of type III. Indeed, it is not possible to distinguish, by means of transformations, between an electric field when $\mathbf{B}=\mathbf{0}$ in the magnetic limit and an electric field in the electric limit.

IV. THE LOCAL REALIZATIONS OF THE MAXWELL GROUP

In this section we study in some detail the local realizations of the magnetic Maxwell group M_m.

The group M_m can be considered as an extension of the Galilei group G by the Abelian group A, being $A=\{([\mathbf{j},\mathbf{k}],\theta,0)\}$ isomorphic to $\mathbf{R}^7$, thus the following sequence is exact

$$1\to A\to M_m\xrightarrow{p}G\to 1. \tag{5.1}$$

The action of M_m on the manifold $X\times F$ is defined by $\overline{g}\ (x,f)=p(\overline{g})(x,f)$, i.e., via the usual Galilean action, where $\overline{g}\equiv([\mathbf{j},\mathbf{k}],\theta,g)\in M_m$, $g\in G$, and $(x,f)\in X\times F$. The isotopy group of a generic point (x,f) is $\overline{\Gamma}_{(x,f)}=p^{-1}(\Gamma_{(x,f)})$, and the corresponding one to the point $f\equiv(\mathbf{B},\mathbf{E})\in F$ is $\overline{\Gamma}_f=p^{-1}(\Gamma_f)$, where $\Gamma_{(x,f)}$ and Γ_f are the isotopy groups defined in Sec. III. A cross-section $\overline{r}$ from $\theta_f\simeq M_m/\overline{\Gamma}_f$ to M_m is defined by $\overline{r}(f)=([\mathbf{0},\mathbf{0}],0,r(f))$, and a section $\overline{s}$ from $\theta_{(x,f)}\simeq M_m/\overline{\Gamma}_{(x,f)}$ to M_m is given by $\overline{s}\ (x,f)=([\mathbf{0},\mathbf{0}],0,s(x,f))$, being r and s the sections defined in (3.8) and (3.9).

The construction of the local representations of the group M_m is made by induction from the finite-dimensional representations of the subgroup $\overline{\Gamma}_{(x_0,f)}$. Among these representations of $\overline{\Gamma}_{(x_0,f)}$ we choose the following ones defined by

$$\mathcal{D}([\mathbf{j},\mathbf{k}],\theta,\gamma)=\exp\{iq[(\mathbf{j}\cdot\hat{\mathbf{B}}+\mathbf{k}\cdot\hat{\mathbf{E}})+m\theta]\}D(\gamma), \tag{5.2}$$

where $\gamma \in \Gamma_{(x_0,f)}$ and D is a matrix representation of $\Gamma_{(x_0,f)}$. The general expression of an induced representation R of M_m by a finite-dimensional representation σ of $\overline{\Gamma}_{(x_0,f)}$ is

$$(R(\overline{g})\psi)(\overline{g}(x,f)) = \sigma(\overline{s}^{-1}(\overline{g}(x,f))\overline{g}\,\overline{s}(x,f))\psi(x,f), \tag{5.3}$$

where $\overline{s}$ is a normalized Borel cross section from $\theta_{(x_0,f)}$ to M_m.

This representation R is a local representation as it is easy to prove if one takes into account that $\sigma(\overline{s}^{-1}(\overline{g}(x,f))\overline{g}\,\overline{s}(x,f))$ has the properties of the gauge matrices (3.3). On the other hand, the choice of the section $\overline{s}$ is not essential because different sections give rise to locally equivalent representations, as one can demonstrate following the theory of locally operating representations.[9] Moreover, if we take equivalent representations σ and σ' of $\overline{\Gamma}_{(x_0,f)}$, they induce locally equivalent representations of M_m. In fact, it can be shown that the induced representations from (5.2) are pseudoequivalent if and only if the representations D of $\Gamma_{(x_0,f)}$ are pseudoequivalent.

If the representation D (of $\Gamma_{(x_0,f)}$) is the restriction to $\Gamma_{(x_0,f)}$ of a matrix representation of the homogeneous Galilei group G_0, then the induced representation R, (5.3), called covariant representation, takes the following expression:

$$\begin{aligned}(R(\overline{g})\psi)(x,f) = \exp\{iq([\mathbf{j},\mathbf{k}]\cdot f - (\mathbf{a}\odot(\mathbf{v},R)_s(t,\mathbf{x}),\\ (b,\mathbf{a})\odot(\mathbf{v},R)(t,\mathbf{x}))\cdot f + im\theta)\}\\ \times D(\mathbf{v},R)\psi(g^{-1}(x,f)).\end{aligned} \tag{5.4}$$

The matrix representation D will be chosen according to the spin of the free physical system.

VI. INVARIANT EQUATIONS

The local representations are not, in general, irreducible representations and they contain a family of irreducible representations labeled by the values of the Casimir operators. These Casimir operators give rise to differential equations for the wave functions ψ that support the local representations. In the following we will consider the cases corresponding to spin 0 and 1/2. We will distinguish among the different types of superorbits of the magnetic limit.

A. Superorbits I and II

The orbital invariants are $\mathbf{B}^2 = C_1 > 0, \mathbf{B}\cdot\mathbf{E} = C_2\ (=0$ on II) and the little-group invariants are given by (4.12′) and (4.13′) for I and (4.12′) and (4.14) for II. We recall that the invariant (4.12′) has orbital character for type II superorbits and in this case the two last terms of the left-hand side vanish.

1. Spin zero

For spin 0 we must take $D(\mathbf{v},R)=1$ in order to obtain the corresponding induced local representation. The expressions for the infinitesimal generators are

$$\begin{aligned}H &= -\partial_t + qi^{\frac{1}{2}}\mathbf{E}\cdot\mathbf{x},\\ \mathbf{P} &= -\nabla + qi^{\frac{1}{2}}\mathbf{B}\wedge\mathbf{x} - qi^{\frac{1}{2}}\mathbf{E}t,\\ J_i &= -\tfrac{1}{2}\epsilon_{ijk}(x_j\partial_k - x_k\partial_j) + J_{ij}(\mathbf{B},\mathbf{E}),\\ K_i &= -t\partial_i + imx_i + K_i(\mathbf{B},\mathbf{E}),\\ \beta_i &= iqB_i, \quad \epsilon_i = iqE_i,\end{aligned} \tag{6.1}$$

where the operators $J_{ij}(\mathbf{B},\mathbf{E})$ and $K_i(\mathbf{B},\mathbf{E})$ act over the functions depending on the variables $(\mathbf{B},\mathbf{E})$ in the following way:

$$\begin{aligned}&[J_i(\mathbf{B},\mathbf{E}),E_j] = \epsilon_{ijk}E_k, \quad [J_i(\mathbf{B},\mathbf{E}),B_j] = \epsilon_{ijk}B_k,\\ &[K_i(\mathbf{B},\mathbf{E}),E_j] = -\epsilon_{ijk}B_k, \quad [K_i(\mathbf{B},\mathbf{E}),B_j] = 0,\\ &[J_i(\mathbf{B},\mathbf{E}),K_j(\mathbf{B},\mathbf{E})] = \epsilon_{ijk}K_k(\mathbf{B},\mathbf{E}).\end{aligned} \tag{6.2}$$

In particular, the restrictions of these operators $J_i(\mathbf{B},\mathbf{E})$ and $K_i(\mathbf{B},\mathbf{E})$ associated to the generators of the isotopy group Γ_f act trivially over $\psi(x;\mathbf{B},\mathbf{E})$. If we replace the generators by their representations in the invariants we get

$$\begin{aligned}&[(i\partial_t + q\phi) - (-i\nabla + q\mathbf{A})^2/2m]\psi(x;\mathbf{B},\mathbf{E})\\ &\quad = \lambda_2\psi(x;\mathbf{B},\mathbf{E}),\end{aligned} \tag{6.3}$$

$$\begin{aligned}&[(i\partial_t + q\phi) - (\mathbf{B}\cdot(-i\nabla + q\mathbf{A}))^2/2m\mathbf{B}^2]\psi(x;\mathbf{B},\mathbf{E})\\ &\quad = \lambda_1\psi(x;\mathbf{B},\mathbf{E}),\end{aligned} \tag{6.4}$$

where $\phi = -\frac{1}{2}\mathbf{E}\cdot\mathbf{x}$ and $\mathbf{A} = \frac{1}{2}(\mathbf{B}\wedge\mathbf{x} - \mathbf{E}t)$ are the components of the four-potential given in Sec. II. Here, we have taken into account the fact that $\mathbf{B}\cdot\mathbf{K}$ and $\mathbf{B}\cdot\mathbf{J}-\mathbf{E}\cdot\mathbf{K}$ are the generators of the isotopy group of $(\mathbf{B},\mathbf{E})$ and, thus, their corresponding terms $J_i(\mathbf{B},\mathbf{E})$ and $K_i(\mathbf{B},\mathbf{E})$ have 0 eigenvalue and, consequently, we can omit them.

Equation (6.3) corresponds to the minimal coupling of a particle of mass m and charge q with an elm potential $(\phi,\mathbf{A})$. Equation (6.4) gives the energy along the direction of $\mathbf{B}$. The addition of the same constant to λ_1 and λ_2 gives a local representation pseudoequivalent to the original one. Its physical interpretation corresponds to a change of the origin of the potential.

Consider a frame where the emf's $\mathbf{B}$ and $\mathbf{E}$ are both parallel to the z axis. Then Eq. (6.4) minus Eq. (6.3) give

$$i(P_x^2 + P_y^2)/2m + (\hat{\mathbf{B}}\cdot\mathbf{J})/m = i\lambda_2 - i\lambda_1 \tag{6.5}$$

or

$$\begin{aligned}&[(-i\partial_x - qA_x)^2 + (-i\partial_y - qA_y)^2]/2m\psi(x;\hat{\mathbf{B}},\hat{\mathbf{E}})\\ &\quad = (\lambda_1 - \lambda_2)\psi(x;\hat{\mathbf{B}},\hat{\mathbf{E}}).\end{aligned} \tag{6.5′}$$

This equation corresponds to the Landau levels for a particle of mass m and charge q in an emf $(\hat{\mathbf{B}},\hat{\mathbf{E}})$ along the z axis. In this case the difference $\lambda_1 - \lambda_2$ is equal to $(\hat{B}/m)\lambda$, with λ integer, since $\exp\{2\pi\mathbf{J}\cdot\hat{\mathbf{B}}\}$ is to be represented by

$+1$. Moreover, the label λ is physically meaningful because there does not exist an operator performing the pseudoequivalence between the realizations associated to different λ's. Note that if $(\mathbf{B},\mathbf{E})\to 0$, Eq. (6.3) becomes the Schrödinger equation for a free particle of mass m and spin 0.

2. Spin 1/2

In this case we need to take the following representation of G_0 [see (5.4)] which determines $\overline{\Gamma}_{(x_0,\mathcal{T})}$:

$$D(\mathbf{v},R)=D_{1/2}(\mathbf{v},R)=\exp\{i\tfrac{1}{2}\boldsymbol{\sigma}\cdot\boldsymbol{\alpha}\}, \tag{6.6}$$

where R is a spatial rotation of angle α around the $\mathbf{n}$ axis, $\boldsymbol{\alpha}=\alpha\mathbf{n}$ and σ_i are the Pauli matrices.

Now, the equation corresponding to the invariant λ_2 is

$$\begin{aligned}&[(i\partial_t+q\phi)-(-i\nabla+q\mathbf{A})^2/2m\\&\quad+(\mathbf{B}\cdot\boldsymbol{\sigma})/2m]\psi(x;\mathbf{B},\mathbf{E})\\&\quad=\lambda_2\psi(x;\mathbf{B},\mathbf{E}).\end{aligned} \tag{6.7}$$

The equation for the other invariant λ_1 remains equal to (6.4) for spin 0. Equation (6.7) contains the correct term of the interaction spin-magnetic field. This equation can be obtained from the Lévy-Leblond[17,18] equation for spin $\frac{1}{2}$ particles with the electromagnetic minimal coupling with four-vector potential $(\phi,\mathbf{A})$.

In the same way as in the case of spin 0, the addition to λ_1 and λ_2 of a constant gives a pseudoequivalent local representation. The difference $\lambda_1-\lambda_2=\lambda$ have discrete values only, which are related to inequivalent local representations associated with the Landau levels for spin $\frac{1}{2}$ particles. The equation of the Landau levels is

$$\begin{aligned}&[(-i\partial_x+qA_x)^2+(-i\partial_y+qA_y)^2]/2m\psi(x;\hat{\mathbf{B}},\hat{\mathbf{E}})\\&\quad=\hat{B}(\lambda\pm 1)/2m\psi(x,\hat{\mathbf{B}},\hat{\mathbf{E}}),\end{aligned} \tag{6.8}$$

where λ is an odd integer, and the choice of ± 1 depends on the spin orientation (up or down).

When $(\mathbf{B},\mathbf{E})\to(\mathbf{0},\mathbf{0})$ Eq. (6.7) becomes the usual Schrödinger equation but, however, it is not possible to obtain by taking this limit the Lévy-Leblond's equation. We had a similar problem when we studied the (relativistic) Maxwell group in I. There we could not get the Dirac equation starting from Casimirs of the Maxwell group and taking the limit $f\Rightarrow 0$. In order to get the Dirac equation (or, here, the Lévy-Leblond equation) it is necessary to look into the problem under a different point of view as it was discussed in the last part of Sec. VI of our preceding paper. About the connection between the Schrödinger equation and the Lévy-Leblond equation see, for instance, Ref. 18.

B. Superorbits III

Here, the orbital invariant is $E^2=C_1>0$ and the little group invariants are given by (4.15). The generators associated to the corresponding local representation have the following expressions:

$$\begin{aligned}H&=-\partial_t+\tfrac{1}{2}i\mathbf{E}\cdot\mathbf{x},\\\mathbf{P}&=-\nabla-\tfrac{1}{2}i\mathbf{E}t,\\J_i&=-\epsilon_{ijk}(x_j\partial_k-x_k\partial_j)+J_i(\mathbf{E})+S_i,\\\mathbf{K}&=-t\nabla+im\mathbf{x}+\mathbf{K}(\mathbf{E}),\\\epsilon&=iq\mathbf{E},\end{aligned} \tag{6.9}$$

where $S_i=0$ or $\frac{1}{2}i\sigma_i$ whether the spin of the particle is 0 or 1/2, respectively. Substituting (6.9) in (4.15) we get the equations

$$\begin{aligned}&[(i\partial_t-q\phi)-(-i\nabla-q\mathbf{A})^2/2m]\psi(x;\mathbf{E})\\&\quad=\lambda'_2\psi(x;\mathbf{E}),\end{aligned} \tag{6.10}$$

$$\tfrac{1}{2}\mathbf{E}\cdot\boldsymbol{\sigma}\psi(x;\mathbf{E})=\lambda'_1\psi(x;\mathbf{E}), \tag{6.11}$$

where $\phi=-\frac{1}{2}\mathbf{E}\cdot\mathbf{x}$ and $\mathbf{A}=-\frac{1}{2}\mathbf{E}t$.

Equation (6.10) corresponds to the minimal coupling of the electric field interacting with the particle of mass m charge q and spin 0 or 1/2. The other equation (6.11) specifies the spin component along the direction of $\mathbf{E}$. In this case, λ'_1 takes only discrete values. The interaction is independent of the spin components and the addition of a constant to λ'_2 is physically irrelevant. The second equation, (6.11), is algebraic and gives no additional information except for reminding us about the spinorial character of ψ.

C. The electric limit

We can make a study for the group M_e similar to the above given for the group M_m concerning local representations and invariant equations. The results are formally equal to those given in Sec. VI B for the superorbits III of the ml, therefore, we will make only brief comment on this point. In Sec. II B the potential $(\phi,\mathbf{A})$ is given by $(-\frac{1}{2}\mathbf{E}\cdot\mathbf{x},-\frac{1}{2}\mathbf{E}t)$ and in the ml $\mathbf{E}=-\nabla\phi-\partial_t\mathbf{A}$. However, by means of a change of gauge we can get $\phi'=-\mathbf{E}\cdot\mathbf{x}$, $\mathbf{A}'=0$, and $\mathbf{E}=-\nabla\phi'$, just as it must be for the el. Thus this situation can be seen at the same time under both points of view: ml or el, there is no difference at all. In conclusion we can say that there is not an interaction properly of electric type, which involves $\mathbf{E}$ as well as $\mathbf{B}$. This is explained if we observe that the operators (∂_t,∇) and the four potential $(\phi,\mathbf{A})$ in the el do not transform in the same way under G_0, and hence cannot be mixed to give interactions, except when $\mathbf{A}=\mathbf{0}$, and in this case we obtain a coincidence with the type III superorbits of the ml.

VII. CONCLUSIONS

With this paper about the nonrelativistic elementary particles interacting with external constant emf's, we have completed a study started in paper I. As in I we have employed a new kind of local realizations of the kinematical group, here the Galilei group, which depend on the

emf, too. The Galilean electromagnetism is very different from the real relativistic electromagnetism. The existence of two limits, magnetic and electric, gives rise to the appearance of two representation groups, one for each limit, both called nonrelativistic Maxwell groups: M_m and M_e. Making use of local representations of the Maxwell groups, we have obtained the invariant equations which describe interacting elementary systems. Some of these equations derived group theoretically, display the electromagnetic minimal coupling.

ACKNOWLEDGMENTS

We would like to thank Professor M. Gadella for helpful technical discussions.

We also wish to thank Comisión Interministerial de Ciencia y Tecnología (CICYT) and Caja de Ahorros de Salamanca for partial financial support.

[1] J. Negro and M. A. del Olmo, J. Math. Phys. **31**, 568 (1990).
[2] E. P. Wigner, Ann. Math. **40**, 149 (1939).
[3] H. Hoogland, Nuovo Cimento B **32**, 427 (1976).
[4] H. Hoogland, "On local and global equivalence of wave functions in quantum mechanics," Ph.D. thesis, Nigmegen University, 1977.
[5] H. Hoogland, J. Phys. A **11**, 1557 (1978).
[6] J. F. Cariñena, M. A. del Olmo, and M. Santander, Physica A **114**, 420 (1982).
[7] M. Asorey, J. F. Cariñena, and M. A. del Olmo, J. Phys. A **16**, 1603 (1983).
[8] J. F. Cariñena, M. A. del Olmo, and M. Santander, J. Phys. A **17**, 3091 (1984).
[9] J. F. Cariñena, M. A. del Olmo, and M. Santander, J. Math. Phys. **26**, 2096 (1985).
[10] O. García-Prada, M. A. del Olmo, and M. Santander, J. Math. Phys. **29**, 1083 (1988).
[11] J. F. Cariñena and M. Santander, J. Math. Phys. **20**, 2168 (1979).
[12] R. Schrader, Fortschr. Phys. **20**, 701 (1972).
[13] M. Le Bellac and J. M. Lévy-Leblond, Nuovo Cimento B **14**, 217 (1973).
[14] H. Bacry, Ph. Combe, and J. L. Richard, Nuovo Cimento A **67**, 267 (1970).
[15] H. Hoogland, J. Phys. **20**, 797 (1978).
[16] J. Beckers and V. Hussin, J. Math. Phys. **24**, 1286 and 1295 (1983).
[17] J. M. Lévy-Leblond, J. Math. Phys. **4**, 776 (1967).
[18] J. F. Cariñena and M. Santander, J. Phys. A **15**, 343 (1982).

Cohomology theory and deformations of Z_2-graded lie algebras

K. C. Tripathy[a)]
International Centre for Theoretical Physics, Trieste, Italy

M. K. Patra
Department of Physics and Astrophysics, University of Delhi, Delhi-110007, India

(Received 27 October 1988; accepted for publication 5 July 1990)

The algebraic cohomology and the spectral sequences for a Z_2-graded Lie algebra are briefly reviewed. The reducibility property of a strongly semisimple Lie superalgebra is established. The role of second and third cohomologies in the deformation of a Lie superalgebra is discussed. Using spectral sequences, the second cohomology of the full BRS algebra is shown to be the ground field and the third cohomology being trivial implies that osp(1,2) is the only graded Lie algebra obtained by deformation of the full BRS algebra. A similar analysis yields the superconformal algebra as a deformation of the super Poincaré algebra. The superconformal algebra so derived contains so(4,1) as the even part, ruling out the existence of negative curvature of a de Sitter universe!

I. INTRODUCTION

The algebraic cohomology (Chevalley cohomology) of Z_2-graded Lie algebras[1] has been generalized to Lie algebras with an arbitrary grading Γ (an Abelian group).[2,3] The Weyl reducibility criteria were examined for a certain class of restricted Z_2-graded Lie algebras.[4] In the present paper, we sharpen some of the earlier results. The Hochschild–Serre method of constructing spectral sequences facilitating the computation of cohomologies of Lie algebras[5] is extended to Z_2-graded Lie algebras (Lie superalgebras). The connection of cohomologies with the deformation (contraction) of Lie superalgebras is systematically analyzed. The deformations of Poisson Lie algebras of complex-valued smooth functions on a symplectic manifold has played a significant role in our factual understanding of quantization of classical systems.[6] The present formulation can be analogously exploited for studying quantization of classical systems. Here, we squarely confine ourselves to the study of deformations of Lie superalgebras and establish its kinship with second and third cohomologies.

This paper is arranged as follows: In Sec. II, we briefly review the algebraic cohomology theory of a Z_2-graded Lie algebra, $L = L_{\bar{0}} + L_{\bar{1}}$. Consider an L_0 module M that is an L_0 module *a fortiori*. Therefore, if L_0 and M are finite dimensional and L_0 is a semisimple Lie algebra, then M is a direct sum of highest weight modules of $L_{\bar{0}}$. Then, under fairly general conditions on M, we prove that the first cohomology group $H^1(L,M) = \{0\}$.

In Sec. III, we prove that for a finite-dimensional strong semisimple Lie superalgebra L, $H^i(L,M) = \{0\}$ for all $i \geqslant 0$ provided $M^L = \{$the L submodule of M annihilated by $L\} = \{0\}$.

The spectral sequence associated with a filtration with respect to a subalgebra of a Lie superalgebra is constructed in Sec. IV, following Hochschild–Serre's work on Lie algebra cohomology.[5] We then make use of spectral sequences to prove a splitting theorem of direct computational interest. Let R be an ideal of L such that L/R is a strongly semisimple Lie superalgebra. Then

$$H^n(L,M) = \sum_{i+j=n} H^i\left(\frac{L}{R},\mathcal{F}\right) \otimes (H^j(R,M))^L, \quad n = 0,1,2,...,$$

where $\mathcal{F}$ is the base field of L considered as a trivial L module and $(H^j(R,M))^L$ is a submodule of $H^j(R,M)$ annihilated by L.

In Sec. V, we briefly review some general results from the theory of deformations of algebraic structures on a Z_2-graded vector space.[6–8] The third cohomology group of L yields the obstructions to the deformation of L. If the deformations of a Lie superalgebra are trivial then L is said to be rigid. The Inönu–Wigner (IW) contraction[9] scheme is extended to the case of Lie superalgebras and the connection between contraction and deformation[8] is established. We also consider a generalization of IW contraction that is peculiar to Lie superalgebras.

In Sec. VI, we consider the application of the ideas developed in previous sections to two cases of physical interest. The first is the case of contraction of $B = \text{osp}(1,2)$ yielding the full BRS algebra A. We compute $H^2(A,A) = \mathcal{F}$ (the base field of A) and $H^3(A,A) = \{0\}$ and therefore conclude that B is the only Lie superalgebra that can be contracted to A. The second example, considered in this section is the super-Poincaré algebra which is obtained by the contraction of the super de Sitter algebra. This is the generalized IW contraction mentioned above. From cohomological computations it is concluded that the super de Sitter algebra is the only algebra that can be contracted to yield the super Poincaré algebra. As a further consequence, we also have that the contraction parameter in this case is identified with the curvature of a de Sitter universe and it must be positive.

II. COHOMOLOGY OF A LIE SUPERALGEBRA

Let $L = L_{\bar{0}} \oplus L_{\bar{1}}$ be a Lie superalgebra. Then, for a,b, and $c \in L$, we have

[a)] Permanent address: Dept. of Physics, and Astrophysics, University of Delhi, Delhi-110007, India.

 0022-2488/90/122822-10$03.00

$$[a,b] = (-1)^{|a||b|}[b,a], \tag{1}$$

$$\sum_{\text{cycl}} (-1)^{|a||c|}[a,[b,c]] = 0. \tag{2}$$

Here, [,] denotes the product in L and the map $a \to |a|$ is the degree map of a homogeneous element $a \in L$. The identity (2) is the graded Jacobi identity. An L module is a Z_2-graded vector space with a bilinear map $L \times M \to M$ written as $(l,m) \to l \cdot m$ for l, l_1, and $l_2 \in L$ and $m \in M$ such that

$$l \cdot M_j \quad M_{j+|l|}, \quad j = \bar{0}, \bar{1}, \tag{3}$$

$$[l_1,l_2] \cdot m = l_1 \cdot (l_2,m) - (-1)^{|l_1||l_2|} l_2 \cdot (l_1 \cdot m). \tag{4}$$

In (3), the sum $j + |l|$ is taken mod 2. Hereafter, whenever the degree occurs, it will be implicitly assumed that it is the Z_2 degree and all sums and products are taken mod 2 unless explicitly stated otherwise.

The tensor powers $T^n V$, $n = 1,2,\ldots$ of a Z_2-graded vector space are Z_2 graded. The Z_2 degree on $T^n V$ is defined as follows

For $i = \bar{0}, \bar{1}$,

$$(T^n V)_i = \{x_1 \otimes \cdots \otimes x_n \in T^n V \,|\, x_1,\ldots,x_n \in V$$

and

$$i = |x_1| + \cdots + |x_n|\}. \tag{5}$$

Hence, the tensor algebra $TV = \Sigma_n T^n V$ is graded by the group $Z \times Z_2$, where Z is the group of integers.

Submodules, ideals, subspaces, etc. of graded algebras and spaces are defined as usual, but are required to be graded with the induced gradation. We also restrict the base field $\mathcal{F}$ to $\mathcal{R}$, the field of reals or $\mathcal{C}$, the complex field. The term "graded" will refer to the Z_2 gradation and other grading groups (e.g., Z) will be explicitly mentioned as and when they appear.

Let V and W be graded vector spaces over $\mathcal{F}$. Let $H(V,W)$ denote the space of morphisms $V \to W$; $H(V,W)$ is a graded vector space with $H(V,W)_{\bar{0}}$ [resp. $H(V,W)_{\bar{1}}$] consisting of all morphisms $V_i \to W_i$ [resp. $V_i \to W_{i+1}$]. The base field $\mathcal{F}$ is trivially graded with $\mathcal{F}_{\bar{0}} = \mathcal{F}$ and $\mathcal{F}_{\bar{1}} = 0$. Then, the dual space $V^* = H(V,\mathcal{F})$ is graded. Similarly, the space $H(V, H(V,\ldots,H(V,W))\cdots)$, where V occurs n times, is graded. Since $H(V,H(V,\ldots H(V,W))\ldots)$ is isomorphic to $T^n(V,W)$, the space of n-linear maps $VX\cdots XV \to W$, $T^n(V,W)$ is also graded.

Let L be a Lie superalgebra and M an L module. An n-linear ($n \geqslant 1$) map $f: LX\cdots XL \to M$ is called graded alternating if

$$f(x_1,\ldots,x_i,x_{i+1},\ldots,x_n) = (-1)^{|x_i||x_{i+1}|} f(x_1,\ldots,x_{i+1},x_i,\ldots,x_n). \tag{6}$$

Let $C^n(L,M) \subset T^n(L,M)$ denote the subspace of n-linear graded alternating maps. By definition $C^0(L,M) = M$ and $C(L,M) = \Sigma_n C^n(L,M)$.

Lemma 1: $C^n(L,M)$ is an L module.

Proof: Let $f \in C^n(L,M)$ and $x \in L$. Then, the module operation $(x,f) \to x \cdot f$ is defined as follows:

$$(x \cdot f)(x_1,\ldots,x_n) = x \cdot f(x_1,\ldots,x_n) - \sum_i k_i f(x,\ldots,[x,x_i]\cdots x_n], \tag{7}$$

where

$$k_i = (-1)^{(|f| + |x_1| + \cdots + |x_i|)(|x|)}.$$

It is easily verified that $x \cdot f \in C^n(L,M)$ and the map $(x,f) \to x \cdot f$ is bilinear. Also, the first condition in (3) is straightforward. To prove the second, let us define $f_x \in C^{n-1}(L,M)$ by

$$f_x(x_1,\ldots,x_{n-1}) = (-1)^{|f||x|} f(x,x_1,\ldots,x_n). \tag{8}$$

Then,

$$(y \cdot f)_x = (-1)^{|y||x|} y \cdot f_x + f_{[x,y]}. \tag{9}$$

From (9), it follows that if the second condition in (3) is true in $C^{n-1}(L,M)$ then, it is true in $C^n(L,M)$. Since $C^0(L,M) = M$ and (3) holds by definition, it is true for all n by induction.

The coboundary operator d: $C^n(L,M) \to C^{n+1}(L,M)$ is an operator on $C(L,M)$ with Z degree 1 and Z_2 degree zero and satisfies the following properties:

$$d(x \cdot f) = x \cdot (df), \tag{10a}$$

$$d(m)(x) = x \cdot m(-1)^{|m||x|}, \tag{10b}$$

and

$$d^2 = 0, \tag{11}$$

where $x \in L$ and $m \in M$. We will define d inductively. It is already defined on $C^0(L,M) = M$ by (10b). Equation (10a) states that d is an L-module homomorphism and (11) is the crucial coboundary property. In the general terminology of homology theory,[10] d is also called a differential.

Definition 1: Let d: $C^{n-1}(L,M) \to C^n(L,M)$ be given. Then, for $x \in L$ and $f \in C^n(L,M)$, we define

$$(df)_x = x \cdot f - d(f_x). \tag{12}$$

Lemma 1: The operator d defined inductively by (12) is a coboundary operator on $C(L,M)$, i.e., d satisfies (10) and (11).

Proof: Let $f \in C^n(L,M)$. The fact that $df \in C^{n+1}(L,M)$ follows from Definition 1 followed by an induction argument.

To prove (10a), we assume its validity for $n > 0$ and show that it is true for $n+1$. For $n = 0$, it is the module condition. Let $g \in C^{n+1}(L,M)$ and $x,y \in L$. Then, from (12) and (9) we have

$$\begin{aligned}
(d(y \cdot f))_x &= x \cdot (y \cdot f) - d((y \cdot f)_x) \\
&= x \cdot (y \cdot f) - (-1)^{|x||y|} d(y \cdot f_x + f_{[y,x]}) \\
&= x \cdot (y \cdot f) - d(f_{[x,y]}) - (-1)^{|x||y|} d(y \cdot f_x) \\
&= ([x,y]) \cdot f - d(f_{[x,y]}) + (-1)^{|x||y|}(y \cdot (x \cdot f) - d(y \cdot f_x)) \\
&= (df)_{[x,y]} + (-1)^{|x||y|}(y \cdot (x \cdot f - d(f_x))) = (df)_{[x,y]} + (-1)^{|x||y|} y \cdot (df)_x = (y \cdot df)_x.
\end{aligned}$$

Since, x is an arbitrary homogeneous element of L we have $y\cdot(df) = d(y\cdot f)$ for all $y\in L$, i.e., d is an L-module homomorphism.

Similarly, assuming the validity of (11) in $C^{n-1}(L,M)$, let $f\in C^n(L,M)$. Then

$$\begin{aligned}(d^2f)_x &= x\cdot(df - d((df)_x))\\ &= x\cdot df - d(x\cdot f + d(f_x))\\ &= x\cdot df - d(x\cdot f) = 0.\end{aligned}$$

Hence, (11) is proved for all n.

The operator d is given explicitly by

$$\begin{aligned}df(x_0,\ldots,x_n) &= \sum k_i x_i\cdot f(x_0,\ldots,\hat{x}_i,\ldots,x_n)\\ &\quad + \sum_{i<j} a_{ij} f([x_i,x_j],x_0,\ldots,\hat{x}_i,\ldots,\hat{x}_j,\ldots,x_n),\end{aligned} \tag{13}$$

where $f\in C^n(L,M)$ and

$$k_i = (-1)^{(|x_0| + \cdots + |x_{i-1}| + |f|)|x_i| + i}$$

$$a_{ij} = (-1)^{(|x_i| + |x_j|)(|x_0| + \cdots + |x_{i-1}|) + |x_j|(|x_{i+1}| + \cdots + |x_{j-1}|) + ij}$$

and the caret "^" in (13) means that the corresponding arguments are to be omitted. In the literature,[1-3] (13) is used to define d. However, we have adopted the inductive definition (12) because of its general nature.

The following is a technical result that will be used in proving certain isomorphisms later.

Lemma 3: Let $f\in C^n(L,M)$. For $0\leqslant j\leqslant n$, let $f_j\in C^j(L,C^{n-j}(L,M))$ be defined by setting

$$f_j(x_1,\ldots,x_j)(x_{j+1},\ldots,x_n) = f(x_1,\ldots,x_n). \tag{14}$$

Then,

$$\begin{aligned}(df)_{j+1}(x_0,\ldots,x_j) &= d(f_j)(x_0,\ldots,x_j)\\ &\quad + (-1)^{j+1} d(f_{j+1}(x_0,\ldots,x_j))\end{aligned} \tag{15}$$

Proof: First, we note that in the first term on the rhs of (15), the underlying module is $C^{n-j}(L,M)$ and in the second term it is M, and the symbol d is used for coboundary opeartors in both. The proof of (15) follows immediately from (13) and (14) and suitable rearrangement of terms.

Definition 2: Let M, N, and P be L modules. Here, M and N are said to be paired to P if there exists a bilinear map $M\times N\to P$, given by $(m,n)\to m\wedge n\in P$ such that, for $x\in L$,

$$x\cdot(m\wedge n) = x\cdot m\wedge n + (-1)^{|m||x|} m\wedge x\cdot n. \tag{16}$$

The pairing of M and N to P induces a pairing of $C^i(L,M)$ and $C^j(L,M)$ to $C^{i+j}(L,P)$. In fact, let $f\in C^i(L,M)$ and $g\in C^j(L,M)$. Then we define $f\wedge g\, C^{i+j}(L,P)$:

$$\begin{aligned}(f\wedge g) = &\sum_{\substack{\sigma(1)<\cdots<\sigma(i)\\ \sigma(i+1)<\cdots<\sigma(i+j)}} a_\sigma f(x_{\sigma(1)},\ldots,x_{\sigma(i)})\\ &\wedge g(x_{\sigma(i+1)},\ldots,x_{\sigma(i+j)}),\end{aligned} \tag{17}$$

where $\sigma\in S_{i+j}$, the permutation group of $i+j$ symbols, σ' is the permutation induced by σ on the subset of odd elements of the set $\{x_1,\ldots,x_{i+j}\}$ and

$$a_\sigma = (-1)^{\sigma+\sigma'+|g|(|x_{\sigma(1)}| + \cdots + x_{\sigma(i)}|)}.$$

It can be checked from (17) that

$$x\cdot(f\wedge g) = x\cdot f\wedge g + (-1)^{|x||f|} f\wedge x\cdot g.$$

Therefore, $(f,g)\to f\wedge g$ is a pairing. We also have

$$(f\wedge g)_x = f_x\wedge g + (-1)^{|f||x|} f\wedge g_x. \tag{18}$$

It follows from (17) and (18) that

$$d(f\wedge g) = df\wedge g + (-1)^i f\wedge dg. \tag{19}$$

The most important example of pairing is the tensor product pairing of M and N to $M\otimes N$, where the module action is defined by (16). We therefore, have a pairing of $C^i(L,M)$ and $C^j(L,N)$ to $C^{i+j}(L,M\otimes N)$.

Remark: In the literature on differential geometry the usual notations are slightly different. The correspondence is as follows.

Let $x\in L$ and $f\in C(L,M)$

$$x\cdot f \sim L_x(f)$$

and

$$f_x \sim i_x(f),$$

where L_x is the Lie derivative with respect to x (now considered as a vector field) and i_x is the interior multiplication by x.

Definition 3: The cohomology groups are defined in the usual way:

$$H^i(L,M) = \frac{\operatorname{Ker} d\cap C^i(L,M)}{\operatorname{Im} d\cap C^i(L,M)} = \frac{Z^i(L,M)}{B^i(L,M)}. \tag{20}$$

Members of $Z^i(L,M)$ are called i cocycles and those of $B^i(L,M)$ i coboundaries. Since d is an L-module homomorphism, $Z^i(L,M)$ and $B^i(L,M)$ are L submodules and hence $H^i(L,M)$ is an L module. It follows from (19) that the pairing of M and N to $M\otimes N$ induces naturally a pairing of $H^i(L,M)$ and $H^j(L,N)$ to $H^{i+j}(L,M\otimes N)$. Therefore, we have an L-module homomorphism:

$$H^i(L,M)\times H^j(L,N)\to H^{i+j}(L,M\otimes N).$$

Let $L = L_{\bar{0}}\oplus L_{\bar{1}}$ be a finite-dimensional Lie superalgebra such that $L_{\bar{0}}$ is a semisimple Lie algebra. Let H be a Cartan subalgebra of $L_{\bar{0}}$. Let M be a finite-dimensional L module and hence, an $L_{\bar{0}}$ module, *a fortiori*. Here $L_{\bar{1}}$ is an $L_{\bar{0}}$ module with respect to the adjoint action. The semisimplicity of $L_{\bar{0}}$ implies that any finite-dimensional $L_{\bar{0}}$ module is a direct sum of irreducible highest weight $L_{\bar{0}}$ modules[11] (the weights being considered relative to a root system for H). Thus suppose $\lambda_1,\ldots,\lambda_k\in H^*$ are the highest weights of the $L_{\bar{0}}$ module $L_{\bar{1}}$. Similarly, let $\mu_1,\ldots,\mu_r$ be the highest weights in M, considered as an $L_{\bar{0}}$ module. If we denote by $V(\nu)$ the highest weight module with highest weight ν then,

$$M = V(\mu_1) + \cdots + V(\mu_r) \tag{21}$$

and

$$L_{\bar{1}} = V(\lambda_1) + \cdots + V(\lambda_k). \tag{22}$$

Now, we can prove the following theorem.

Theorem 1: Let L be a Lie superalgebra such that $L_{\bar{0}}$ is semisimple. Let M be a finite-dimensional L module. Let $\{\mu_j; j = 1,\ldots,r\}$ be the highest weights in M (as an $L_{\bar{0}}$ mod-

ule), with respect to a Cartan subalgebra H of $L_{\bar{0}}$. Let $\{\lambda_i : i = 1,...,k\}$ be the highest weights in $L_{\bar{1}}$ considered as an $L_{\bar{0}}$ module. Suppose, $\lambda_i \neq \mu_j$. Then, $H^1(L,M) = \{0\}$.

Proof: Let $L_{\bar{0}} = H + L_{\bar{0}}^+ + L_{\bar{0}}^-$ be the decomposition of $L_{\bar{0}}$, relative to H, where $L_{\bar{0}}^+$ (resp. $L_{\bar{0}}^-$) is the subalgebra generated by $+ve$ (resp. $-ve$) root vectors. Then $x \in L_{\bar{0}}^+$ annihilates the maximal vectors.

Suppose f is a one-cocycle. Then

$$df(x,y) = 0$$
$$\Rightarrow f([x,y]) = (-1)^{|x||f|} x \cdot f(y) - (-1)^{|y|(|f|+|x|)} y \cdot f(x). \tag{23}$$

We assume first that f is even, i.e., $f \in C^1(L,M)_{\bar{0}}$. Then, for $x,y \in L_{\bar{0}}$, (23) reduces to $f([x,y]) = x \cdot f(y) - y \cdot f(x)$. Therefore by Whitehead's first lemma[11] there exists $m \in M$ such that $f(x) = x \cdot m$, for all $x \in L_{\bar{0}}$. Let $u_{\lambda_i} \in V(\lambda_i)$ be a maximal vector. Then $x \cdot u_{\lambda_i} = 0$, if $x \in L_{\bar{0}}^+$. Now, we put u_{λ_i} for y in (17). Thus, we obtain

$$x \cdot f(u_{\lambda_i}) - u_{\lambda_i} f(x) = 0, \quad \text{for all } x \in L_{\bar{0}}^+$$
$$\Rightarrow x \cdot f(u_\lambda) - u_{\lambda_i} \cdot x \cdot m = 0, \quad \text{since } x \cdot u_{\lambda_i} = [x, u_{\lambda_i}] = 0,$$

or

$$x \cdot [f(u_{\lambda_i}) - u_{\lambda_i} \cdot m] = 0.$$

Hence, $L_{\bar{0}}^+$ annihilates the vector $(f(u_{\lambda_i}) - u_{\lambda_i} \cdot m) \in M$. Moreover, if $h \in H$, then

$$f([h,u_{\lambda_i}]) = \lambda_i(h) \cdot f(u_{\lambda_i})$$
$$= h \cdot f(u_{\lambda_i}) + \lambda_i(h) u_{\lambda_i} \cdot m - h \cdot u_{\lambda_i} \cdot m$$
$$\Rightarrow h \cdot (f(u_{\lambda_i}) - u_{\lambda_i} \cdot m)$$
$$= \lambda_i(h) \cdot (f(u_{\lambda_i}) - u_{\lambda_i} \cdot m).$$

Hence, $f(u_{\lambda_i}) - u_{\lambda_i} \cdot m$ is a maximal vector in M with weight λ_i. But, we have assumed that $\lambda_i \neq \mu_j$ and the latter exhausts the weights of the maximal vectors in M. Therefore, $f(u_{\lambda_i}) = u_{\lambda_i} \cdot m$.

Next, let α be a simple root of $L_{\bar{0}}$ relative to some base in the root system relative to H and let $y_{-\alpha} \in L_{\bar{0}}^-$. Then, $[y_{-\alpha}, u_{\lambda_i}] \in V_{\lambda_i - \alpha}$, the weight space corresponding to the weight $\lambda_i - \alpha$ in $V(\lambda_i)$.

Using $u_{\lambda_i - \alpha} = [y_{-\alpha}, u_{\lambda_i}]$, we have

$$f(u_{\lambda_i - \alpha}) = y_{-\alpha} \cdot f(u_{\lambda_i}) - u_{\lambda_i} \cdot y_{-\alpha}(m)$$
$$= y_{-\alpha} \cdot f(u_{\lambda_i}) - u_{\lambda_i} \cdot y_{-\alpha} \cdot (m)$$
$$= [y_{-\alpha}, u_{\lambda_i}] \cdot m = u_{\lambda_i - \alpha} \cdot m. \tag{24}$$

Since the vectors $(\mathrm{ad}\, y_{-\alpha_i})^{s_i} \cdots (\mathrm{ad}\, y_{-\alpha_1})^{s_1} \cdot u_{\lambda_i}$, with α_i simple roots and $s_i \geq 0$ generate the submodule $V(\lambda_i)$ of $L_{\bar{1}}$, it is seen from (24) that by induction $f(u) = u \cdot m$ for all $u \in V(\lambda_i)$. But $L_{\bar{1}} = \Sigma_i V(\lambda_i)$. Hence, $f(u) = u \cdot m$ for all $u \in L_{\bar{1}}$. As it is already known that $f(x) = x \cdot m$, for all $x \in L_{\bar{0}}$, we see that every even one-cocycle is a coboundary.

Now, let g be an odd one-cocycle. The cocycle condition for $x \in L_{\bar{0}}$ and $u \in L_{\bar{1}}$ is

$$g([x,u]) = x \cdot g(u) + u \cdot g(x) \tag{25a}$$

and for $x,y \in L_{\bar{0}}$,

$$g([x,y]) = x \cdot g(y) - y \cdot g(x). \tag{25b}$$

Using an argument as above we arrive at the conclusion that there is an element $m_1 \in M_{\bar{1}}$ such that $g(x) = x \cdot m_1$, for all $x \in L_{\bar{0}}$ and $g(u) = -u \cdot m_{\bar{1}}$, $u \in L_{\bar{1}}$. But, these are precisely the coboundary relations for odd elements of $C^1(L,M)$. Hence, we conclude that $H^1(L,M) = \{0\}$.

Corollary: Let L and M be as in the theorem. Suppose $L_{\bar{1}}$ is an irreducible $L_{\bar{0}}$ module with highest weight λ. Then, $H^1(L,M) = \{0\}$, if none of the highest weights in M (as an $L_{\bar{0}}$ module) equals λ. In particular, for the adjoint representation ($M = L$), $H^1(L,L) = \{0\}$, provided $[L_{\bar{1}}, L_{\bar{1}}] \neq 0$ and the highest weight in $L_{\bar{1}}$ is not a maximal root in $L_{\bar{0}}$.

Proof: The first assertion of the corollary follows directly from the theorem. The second is proved as follows.

In the proof of Theorem 1, it was shown that if f is an even one-cocycle then $f(u_\lambda) - u_\lambda \cdot m$ must be a maximal vector in $M_{\bar{1}}$ ($= L_{\bar{1}}$ in this case). For the adjoint representation we have $m \in L_{\bar{0}}$ such that for all $x \in L_{\bar{0}}$, $f(x) = [x,m]$. Since, $L_{\bar{1}}$ is an irreducible $L_{\bar{0}}$ module, u_λ (up to a scalar multiple) is the only maximal vector. Therefore, $f(u_\lambda) - u_\lambda \cdot m = c u_\lambda$, for some scalar c.

For any $u,v \in L_{\bar{1}}$ the cocyle condition gives

$$f([u,v]) = u \cdot f(v) + v \cdot f(u),$$

or

$$[u,v] \cdot m = u \cdot f(v) + v \cdot f(u), \because [u,v] \in L_{\bar{0}}$$
$$\Rightarrow [u \cdot m, v] + [u, v \cdot m] = u \cdot f(v) + v \cdot f(u). \tag{26}$$

The last relation follows from the Jacobi identity and the fact that we are considering the adjoint representation. Moreover, applying ad $y_{-\alpha_i}$ (α_i a simple root) to u_λ, we get from (23)

$$f(y_{-\alpha_i} \cdot u_\lambda) = f([y_{-\alpha_i}, u_\lambda])$$
$$= y_{-\alpha_i} \cdot f(u_\lambda) - u_\lambda \cdot f(y_{-\alpha_i})$$
$$= y_{-\alpha_i} \cdot (u_\lambda \cdot m + c u_\lambda) - u \cdot y_{-\alpha_i} \cdot m$$
$$= [y_{-\alpha_i}, u_\lambda] \cdot m + c [y_{-\alpha_i}, u_\lambda].$$

Successive applications of ad $y_{-\alpha_i}$'s and the fact that they generate $L_{\bar{1}}$ lead to the conclusion that, for all $u \in L_{\bar{1}}$,

$$f(u) = u \cdot m + cu. \tag{27}$$

From (26) and (27) one gets

$$[u \cdot m, v] + [u, v \cdot m] = [u \cdot m, v] + [u, v \cdot m] + 2c[u,v]$$

i.e., $2c[u,v] = 0$, for all $u,v, \in L_{\bar{1}}$.

Since $[L_{\bar{1}}, L_{\bar{1}}] \neq 0$ by hypothesis, we conclude that $c = 0$. Hence, every even one-cocycle is a coboundary and $H^1(L,L)_{\bar{0}} = \{0\}$.

Next, let g be an odd one-cocycle,

i.e., $g: L_{\bar{0}} \to L_{\bar{1}}$ and $L_{\bar{1}} \to L_{\bar{0}}$.

Since it is given that the highest weight in $L_{\bar{1}}$ is not a maximal root (= highest weight in the adjoint representation $L_{\bar{0}}$ on itself) we conclude from the theorem that $H^1(L,L)_{\bar{1}} = \{0\}$ and the corollary is proved.

Remark: We recall the fact that in the adjoint representation a one-cocycle is a derivation of L and a one-coboundary is an inner derivation. Then we have an immediate application of the preceding corollary: For the algebras $B(m,n)$,

$D(m,n)$, $F(4)$, and $G(3)$ (see Ref. 12 for the classification scheme of Lie superalgebras) all derivations are inner since these algebras satisfy the conditions in the corollary.

III. HIGHER-ORDER COHOMOLOGIES

It is well-known that $H^n(G,V) = \{0\}$, for all $n \geqslant 0$, G a semisimple Lie algebra and V, a nontrivial G module.[11,13] In the case of Lie superalgebras, however, we have to impose the condition of strong semisimplicity to ensure the triviality of the cohomology modules.

Definition: A Lie superalgebra L is defined to be strongly semisimple if all its modules are semisimple, i.e., all the nontrivial representations of L are completely reducible. We then have the following theorem.

Theorem 2: Let L be a finite-dimensional strongly semisimple Lie superalgebra and M a nontrivial finite-dimensional L module. Then, $H^n(L,M) = \{0\}$.

Proof: The module M can be split up into a direct sum of irreducible L modules and the projection maps $\Pi_i : M \to M_i$, where M_i is an irreducible summand, commute with d. Thus, we can assume without loss of generality that M is an irreducible L module.

In the context of representation theory, let $\phi : L \to \text{End } M$ be defined by $x \cdot m = \phi(x) \cdot m$, then ϕ is an algebra homomorphism. Since M is nontrivial, $Q = \text{Ker } \phi$ is either zero or a proper ideal of L. Let R be an ideal of L complementary to Q. The restriction of ϕ to R is an isomorphism. Then, we can work inside R and the arguments that follow hold good for $O \subset R \subset L$. Therefore, one can assume that ϕ is faithful.

We next recall the definition of supertrace form. Let $u = \left(\begin{smallmatrix} a & e \\ f & b \end{smallmatrix}\right)$ be an $(m+n) \times (m+n)$ matrix, where a is $m \times m$ and b an $n \times n$ matrix. The supertrace of u is defined as $\text{str } u = \text{tr } a - \text{tr } b$. Now, let $\beta : L \times L \to \mathcal{F}$ be the bilinear form defined as $\beta(x,y) = \text{str}(\phi(x) \cdot \phi(y))$. We note that β is an even, supersymmetric, and associative bilinear form on L. Now, suppose L is strongly semisimple. Then, according to a theorem of Djokovic and Hochschild[14] L is a direct sum of semisimple Lie algebras and Lie superalgebras in the series $B(0,n)$ $[= \text{osp}(1,2n)]$. Therefore, from the structure of $B(0,n)$ it follows that β is nondegenerate, i.e., $\text{Ker } \beta = \{x \in L \mid \beta(x,y) = 0 \forall x \in L\} = 0$. Let $m = \dim L_{\bar{0}}$ and $n = \dim L_{\bar{1}}$. Let $\{u_i\}_{i=1}^m$ and $\{v_i\}_{i=1}^n$ be bases in $L_{\bar{0}}$ and $L_{\bar{1}}$, respectively, and $\{u^i\}$ and $\{v^i\}$ the corresponding dual bases with respect to β. Define

$$C = \sum_{i=1}^{m} \phi(u_i)\phi(u^i) - \sum_{j=1}^{n} \phi(v_j) \cdot \phi(v^j). \tag{28}$$

Then, $C \in \text{End } M$ commutes with $\phi(x)$, for all $x \in L$. Here, C is a Casimir invariant. Since M is irreducible $C = cI$, c being some scalar (Schur lemma). But,

$$\text{Str } C = \sum \text{tr}(\phi(u_i)\phi(u^i)) - \sum \text{tr}(\phi(v_j)\phi(v^j))$$

$$= \sum \beta(u_i,u^i) - \sum \beta(v_j,v^j) = m - n.$$

Also, $\text{str } C = \text{str}(cI) = c(p-q)$, where $p = \dim M_{\bar{0}}$ and $q = \dim M_{\bar{1}}$. From the structure of the strongly semisimple Lie superalgebras[12] it is known that $m \neq n$. Hence, $c \neq 0$ and thus C is invertible. Therefore, we can show as in the case of semisimple Lie algebras that for any n-cocycle f, $Cf = dg$, where $g \in C^{n-1}(L,M)$. Since c is invertible $f = d(C^{-1}g)$ and hence $H^n(L,M) = \{0\}$, $n = 1,2,\ldots$.

IV. SPECTRAL SEQUENCES

In this section we briefly consider the extension of the Hochschild–Serre method[5] of constructing spectral sequences by introducing a filtration with respect to a subalgebra of a Lie algebra to the case of Lie superalgebras.

Let S be a subalgebra of L. Let $A^n = C^n(L,M)$, $A = \Sigma_n A^n$, and d be the coboundary operator. Then, $\{A,d\}$ is a Z-graded differential L module.[10] A filtration on A is a family of submodules $\{A_j\}$ such that

(1) $\cdots \supset A_j \supset A_{j+1} \cdots$

and

(2) $\bigcup_j A_j = A.$ (29)

If the differential d preserves the filtration, then A is said to be a filtered, graded differential module. A particular filtration with respect to the subalgebra is defined as follows:

$$A_j = 0 \quad \text{if } j > n,$$
$$= A \quad \text{if } j \leqslant 0,$$

and for $0 < j \leqslant n$, $A^n \cap A_j$ consist of all n cochains that vanish if $n - j + 1$ arguments belong to S. We thus have a bounded, descending filtration $A = A_0 \supset A_1 \cdots \supset A_j \supset A_{j+1} \supset \cdots$. Each submodule has intrinsic Z_2-grading besides the usual Z grading. Let us denote $A^{j,i} = A^{i+j} \cap A_j$. The filtration, defined above, gives rise to a spectral sequence,[10] i.e., a family $\{E_r,d_r\}$ of Z-bigraded modules, with d_r having bidegree $(r, -r+1)$ and such that the cohomology module,

$$H(E_r) = E_{r+1}. \tag{30}$$

The differential operator d_r $(d_r^2 = 0)$ is induced by d. Omitting the details we mention the following well-known result.[10]

Theorem 3: Let

$$Z_r^{p,q} = \{x \in A^{p,q} \mid dx \in A^{p+r,q-r+1}\},$$

$$B_r^{p,q} = \{x \in A^{p,q} \mid x = dy \text{ for some } y \in A^{p-r+1,q+r}\},$$

and

$$E_r^{p,q} = (Z_r^{p,q} + A_{p+1})/B_r^{p,q} + A_{p+1}. \tag{31}$$

Further, let d_r be the differential operator induced by d on

$$E_r = \sum_{p,q} E_r^{p,q}.$$

Then $\{E_r,d_r\}$ is a spectral sequence. Further, we have the following identifications.

Let

$$E_r^p = \sum_q E_r^{p,q},$$

$$E_0^p \simeq A_p/A_{p+1},$$

$$E_1^p \simeq H(A_p/A_{p+1}) = H(E_0^p).$$

We also have $E_2^{p,q} \to E_\infty^{p,q}$, i.e., there exist natural isomorphisms

$$E_{\infty}^{p} \cong (H(A))_p/(H(A))_{p+1}. \tag{32}$$

In (32), $(H(A))_p \subset H(A)$ is given by the filtration induced on $H(A)$ by the filtration on A.

We next note that $x \cdot A_p \subset A_p$ for $x \in S$. The submodule $A^{p,0} = A_p \cap A^p$ is the submodule of all p-linear maps $f: L \to M$ which vanish if one of the arguments of f belongs to S. We can therefore identify $A^{p,0} = C^p(L/S,M)$, where L/S is a Z_2-graded vector space. Hence, $C^p(L/S,M)$ is an S module.

Let $s'_{p,q}: A^{p,q} \to C^q(S, C^p(L/S,M))$ be defined by setting

$$s'_{p,q}(f)(x_1,...,x_q)(y^*_{q+1},...,y^*_{p+q}) = f(x_1,...,x_q,y_{q+1},...,y_{p+q}), \tag{33}$$

where $x_1,...,x_q \in S$ and $y^*_{q+1},...,y^*_{p+q}$ are the respective images of $y_{q+1},...,y_{p+q} \in L$ under the canonical projection $L \to L/S$. The function $s'_{p,q}$ is well-defined since (33) is independent of the choice of $y_{q+1},...,y_{p+q}$. This is because f vanishes if any of the $y_{q+1},...,y_{p+q}$ belong to S. Moreover, $A_{p+1} \subset \operatorname{Ker} s'_{p+q}$ and hence we have a family $\{s_{p,q}\}$ of mappings

$$s_{p,q}: A^{p,q}/A^{p+1,q} \to C^q(S, C^p(L/S,M)). \tag{34}$$

But, by definition, $E_0^{p,q} = A^{p,q}/A^{p+1,q}$: Hence there exists a map $\phi: E_0 \to C(S, C(L/S,M))$. We can now show that ϕ is an isomorphism.

Let d_s and d_0 denote the coboundary operator on S and E_0, respectively.

Theorem 4: There is an isomorphism ϕ between the Z-bigraded modules E_0 and $C(S, C(L/S,M))$ such that $\phi \cdot d_0 = d_s \cdot \phi$.

The fact that ϕ is an isomorphism can be shown by modifying the proof in Ref. 5 to take care of the intrinsic Z_2 degree. The relation $\phi \cdot d_0 = d_s \cdot \phi$ follows from Lemma 3 and the definition of ϕ.

Corollary: ϕ induces an isomorphism $\phi_1 = \{\phi_1^{p,q}\}$ of each homogeneous component $\phi_1^{p,q}: E_1^{p,q} \overset{\text{onto}}{\to} H^q(S, C^p(L/S,M))$.

From the above theorem and its corollary we get the following identifications:

$$E_0^{p,0} \simeq C^p(L/S,M)$$

and

$$E_1^{p,0} \simeq (C^p(L/S,M))^S,$$

where the rhs is the subspace of $C^p(L/S,M)$, annihilated by S. Suppose $S = L_{\bar{0}}$. Then, $L/S = L_{\bar{1}}$ and $E_0^{p,0} \simeq C^p(L_{\bar{1}},M)$ and $E_1^{p,0} \simeq (C^p(L_{\bar{1}},M))^{L_{\bar{0}}}$.

Next, we suppose that S is an ideal. Then, $H^n(S,M)$ is an L/S module. Let $d_{L/S}$ denote corresponding coboundary operator. There is an isomorphism τ: $H^p(S, C^q(L/S,M)) \overset{\text{onto}}{\to} C^q(L/S, H^p(S,M))$. Indeed, τ is constructed as follows.

Let τ' denote the vector space isomorphism from $C^p(S, C^q(L/S,M)) \overset{\text{onto}}{\to} C^q(L/S, C^p(S,M))$ defined by

$$\tau'(f)(x^*_{p+1},...,x^*_{p+q})(y_1,...,y_p) = a(f(y_1,...,y_p))(x^*_{p+1},...,x^*_{p+q}), \tag{35}$$

where

$$a = (-1)^{(|y_1| + \cdots + |y_p|)(|x^*_{p+1}| + \cdots + |x^*_{p+q}|)},$$

$y_1,...,y_p \in S$ and $x^*_{p+1},...,x^*_{p+q} \in L/S$.

Further, we have the identity

$$\tau'(d_s f)(x^*_{p+1},...,x^*_{p+q}) = d_{L/S}(\tau^*(f)(x^*_{p+1},...,x^*_{p+q})),$$

which is a consequence of Lemma 3. Hence, τ' induces the required isomorphism τ on the corresponding cohomology modules.

Theorem 5: Let s be an ideal in L. There is an isomorphism $\psi_1 = \{\psi_1^{p,q}\}$ of E_1 onto $C(L/S, H(S,M))$ with

$$\psi_1^{p,q}: E_1^{p,q} \to C^p(L/S, H^q(S,M)).$$

If d_1 denotes the coboundary operator on E_1 then,

$$\psi_1^{p,q} \cdot d_1 = (-1)^q d_{L/S} \cdot \psi_1^{p,q}. \tag{36}$$

Proof: Let $\psi_1^{p,q} = \tau \cdot \phi_1^{p,q}$ where $\phi_1 = \{\phi_1^{p,q}\}$ is the isomorphism given in the corollary to Theorem 4. To prove (36), let $f \in A_p \cap A^{p+q}$ be a representative of an element in $E_1^{p,q}$. Then $df \in A_{p+1}$. From Lemma 3 we have

$$(df)_{p+1}(x_0,...,x_p) = d(f_p)(x_0,...,x_p) + (-1)^{p+1} \cdot d(f_{p+1}(x_0,...,x_p)),$$

where $x_0,...,x_p \in S$. But then $f_{p+1}(x_0,...,x_p) = 0$. Further, we have, on passing to the quotients

$$f_p = (-1)^{pq} \tau(f) \quad \text{and} \quad (df)_{p+1} = (-1)^{(p+1)q} \cdot \tau(df). \tag{37}$$

Equation (36) follows from (37) and Theorem 1.

Corollary: There is an isomorphism $\psi_2 = \{\psi_2^{p,q}\}$ from E_2 onto $H^p(L/S, H^q(S,M))$.

We now suppose that S is an ideal in L such that L/S is strongly semisimple. From Theorem 2, $H^2(L/S,P) = \{0\}$ for any L/S module P. Therefore, there exists a subalgebra $K \subset L$ such that $L = K +)S$, where $K \simeq L/S$ (Levi decomposition).[11]

Theorem 6: Let L, S, and K be as above. Then

$$H^n(L,M) \simeq \sum_{p+q=n} H^p(L/S, \mathcal{F}) \otimes (H^q(S,M)^L).$$

The proof of the above theorem is in several steps, we prove first the following lemmas.

Lemma 4:

$$H^p(L/S, (H^q(S,M))^K) \simeq H^p(K, \mathcal{F}) \otimes (H^q(L/K,M))^K.$$

We note that $H^p(L/S, (H^q(S,M))^K) \simeq H^p(K, (H^q(S,M))^K)$. Now, let $R = (H^q(S,M))^K$. The K operators are zero on R. Moreover, we can identify $C^p(K,R) \simeq C^p(K, \mathcal{F}) \otimes R$. But since K acts trivially on R, $H^p(K,R) \simeq H^p(K, \mathcal{F}) \otimes R$. Hence, we prove the lemma if we can show that $R = (H^q(S,M))^K \simeq (H^q(L/K,M))^K$. This is done as follows: $H(S,M)$ is an L module annihilated by S. Therefore, $(H(S,M))^K \simeq (H(S,M))^L$. Now, we have a K-module decomposition $C^q(S,M) = d(C^{q-1}(S,M) \oplus U)$. Therefore, if the cohomology class of f belongs to $(H^q(S,M))^L$ then $x \cdot f \in d(C^{q-1}(S,M))$, for all $x \in K$. But U is a K submodule. Hence, $x \cdot f = 0$. Now, let $C^{q-1}(S,M) = Z^{q-1}(S,M) \oplus V$ be a K-module decomposition. Then, writing $g = h + k \in C^{q-1}(S,M)$, where $k \in V$ and $h \in Z^{q-1}(S,M)$, suppose $dg \in (C^q(S,M))^K$. Then

$x \cdot g \in Z^{q-1}(S,M)$, for all $x \in K$. We thus have $x \cdot k = 0$. Moreover, $dg = dk$ and hence $B^q(S,M) = d((C^{q-1}(S,M))^K)$. Now, we can identify $(C^q(S,M))^K$ with $C^q(L/K,M)$ in a natural fashion and we have just shown that

$$(H^q(S,M))^K = \frac{(C^q(S,M))^K \cap Z^q(S,M)}{d((C^{q-1}(S,M))^K)}.$$

Thus $H^q(S,M) = (H^q(L/K,M))^K$ and the lemma is proved.

Lemma 5: There exists a homomorphism

$$\zeta: H(K,\mathcal{F}) \otimes (H(L/K,M))^K \text{ into } H(L,M).$$

Proof: First, we prove that the restriction homomorphism $H(L,\mathcal{F}) \to H(K,\mathcal{F})$ is onto. Let π be the projection homorphism of L onto K, corresponding to the decomposition $L = K +)\ S$ and let $f \in Z^n(K,\mathcal{F})$. Let $g \in Z^n(L,M)$ be defined as $g(x_1,\ldots,x_n) = f(\pi(x_1),\ldots,\pi(x_n))$. The restriction of g to K coincides with f. Now, since K is strongly semisimple $H(K,\mathcal{F})$ is isomorphic to the exterior algebra over a subspace of $H(K,\mathcal{F})$ spanned by primitive elements. This fact can be demonstrated as in the case of ordinary Lie algebras.[15] This, together with fact that the restriction homomorphism is onto implies that there is an injective graded algebra homomorphism $\sigma: H(K,\mathcal{F}) \to H(L,\mathcal{F})$ inverse to the restriction homomorphism. Let ν denote the natural homomorphism $(H(L/K,M))^K \to H(L,M)$ and let Λ denote the pairing of $H(L,\mathcal{F})$ and $H(L,M)$ to $H(L,M)$ induced by the tensor product pairing of $\mathcal{F}$ and M to M. Let ζ denote the map $H(K,\mathcal{F}) \otimes H(L/K,M)^K \to H(L,M)$ defined by $\zeta(u \otimes v) = \sigma(u) \wedge \nu(v)$.

Proof of Theorem 6: The homomorphism ζ in Lemma 5 is an isomorphism onto $H(L,M)$. This is seen as follows.

From Theorem 5, and Lemma 4, we have an isomorphism $\psi: E_2^{p,q} \to H^p(L/S,H^q(S,M)) \simeq H^p(K,\mathcal{F}) \times (H^q(L/K,M))^K$. Let f be a p cocycle for L in $\mathcal{F}$ which belongs to the cohomology class $\sigma(u)$, where $u \in H^p(K,\mathcal{F})$ and ζ is the map defined in Lemma 5. Similarly, let g be a q cocycle in the class $\nu(v)$. Then, $f \wedge g$ is a $(p+q)$ cocycle belonging to the cohomology class of $\zeta(u \otimes v)$. Therefore, $f \wedge g$ determines an element h in $E_2^{p,q}$ such that $\psi(h) = u \otimes v$. Since ψ is an isomorphism onto, $E_2^{p,q}$ consists of sums of elements such as h above. Thus every elment of $E_2^{p,q}$ has a representative in $A^{p,q}$ which is a cocycle. The coboundary operator d_2 on E_2 is induced by d and since $dh = 0$, we conclude that d_2 is zero on E_2. Consequently, $E_3 \simeq H(E_2) = E_2$ and every element of E_3 is represented by a cocycle. Hence, by the same arguments we conclude $E_2 \simeq E_3 \simeq \cdots E_\infty$, i.e., $E^{p,q} \simeq E_2^{p,q}$ and $f \wedge g$ is a representative cocycle of $h \in E^{p,q}$ under this correspondence.

Also, $H^n(L,M)$ is a finite-dimensional space and therefore, from (32) we get

$$H^n(L,M) \simeq \sum_{p+q=n} E_\infty^{p,q},$$

where the sum is a vector space direct sum. Moreover, by choosing a suitable basis for $H^n(L,M)$ we get linear isomorphism $\alpha: H(L,M) \xrightarrow{\text{onto}} E$ which is identity on $(H^n(L,M))_p / (H^{n+1}(L,M))_{p+1}$ and $\alpha(\zeta(u \otimes v))$ is the element in $E_\infty^{p,q}$ that corresponds to h. Hence, $\alpha \cdot \zeta$ is an isomorphism onto E_∞. Since α is a linear isomorphism ζ is an isomorphism. Theorem 6 is an immediate consequence of this fact.

V. DEFORMATION AND CONTRACTION

In this section, Gerstenhaber's theory of deformation of algebraic structures[7] on a given vector space is extended to cover Z_2-graded structures. Let $L = L_{\bar{0}} \oplus L_{\bar{1}}$ be a finite-dimensional Lie superalgebra over a field $\mathcal{F}$ and let V be the underlying vector space, i.e., $L = \{V,[\ ,\ ]\}$. Let $P[\lambda]$ be the polynomial ring over $\mathcal{F}$ in the indeterminate λ. Let k denote the field of fractions of $P[\lambda]$. Let $V_k = V \otimes k$ be the extension of V. Any bilinear map $V \times V \to V$ can be uniquely extended to a bilinear map $V_k \times V_k \to V_k$. Suppose there is an even[16] bilinear map $f: V_k \times V_k \to V_k$ expressible in the form

$$f_\lambda(a,b) = \sum_{r=0} \lambda^r f_r(a,b)$$

$$= [a,b] + \sum_{r=1} \lambda^r f_r(a,b), \tag{38}$$

where each $f_r: V \times V \to V$ is an even bilinear map. Suppose further that $(a,b) \to f_\lambda(a,b)$ define a Lie superalgebra structure on V_k. Then, we say that $\{V_k, f_\lambda\}$ is a one-parameter family of deformations of L. The bilinear map f_λ must satisfy (1) and (2). In terms of coefficients f_r, we have

$$f_r(a,b) = (-1)^{|a||b|} f_r(b,a), \tag{39}$$

$$\sum_{\text{cycl.}} \sum_{r+s=n} (-1)^{|a||c|} f_r(f_s(a,b),c) = 0. \tag{40}$$

Equation (39) states that $f_r \in C^2(L,L)_{\bar{0}}$, where L acts on itself via the adjoint action. The conditions (40) are known as integrability conditions. For $n = 1$, (40) reduces to

$$\sum_{\text{cycl.}} (-1)^{|a||c|} \{f_1([a,b],c) + [f_1(a,b),c] = 0. \tag{41}$$

But, the left-hand side of the above equation is equal to $-df_1$, where d is the coboundary operator on $C(L,L)$. Hence, the first $(n = 1)$ integrability condition states that f_1 must be a two-cocycle. An element $f \in Z^2(L,L)$ is said to be integrable if it is the first term f_1 of a one-parameter deformation series (38). Putting $n = 2$ in (36) we get,

$$\sum_{\text{cycl.}} (-1)^{|a||c|} \{f_2([a,b],c) + [f_1(a,b),c]\}$$

$$= \sum_{\text{cycl.}} (-1)^{|a||c|} f_1(f_1(a,b),c))$$

or

$$-df_2(a,b,c) = \sum (-1)^{|a||c|} f_1(f_1(a,b),c)). \tag{42}$$

The rhs of (42) is a three-cocycle if f_1 is a two-cocycle and the condition requires that this three-cocycle must be a coboundary. In general, we have

$$df_n(a,b,c) = \sum_{\text{cycl.}} \sum_{\substack{i+j=n \\ i,j<n}} (-1)^{|a||c|} f_i(f_j(a,b),c). \tag{43}$$

We can show that the rhs of (43) is a three-cocycle provided

the f_i's satisfy the integrability condition of $i<n$. This is done by appropriate modification of Gerstenhaber's proof,[6,7] to take care of the Z_2 grading. Therefore, we conclude that if $H^3(L,L)=\{0\}$, then any two cocycle is integrable. However, this condition is not necessary. The cohomology class of the three-cocycle in the rhs of (42) is called the first obstruction to integrability.

Definition: Let $\{f_\lambda\}$ and $\{g_\lambda\}$ be two deformations of L. Let

$$\phi_\lambda = 1 + \sum \lambda^s \phi_s \tag{44}$$

be an invertible element in End (V_k), such that $\phi_s \in \mathrm{End}(v)$. The deformations $\{f_\lambda\}$ and $\{g_\lambda\}$ are said to be equivalent if, for $a,b,c \in V$,

$$f_\lambda(a,b) = \phi_\lambda^{-1} \cdot g_\lambda(\phi_\lambda(a),\phi_\lambda(b)) \tag{45}$$

A family of deformation $\{g_\lambda\}$ is defined to be trivial if it is equivalent to the identity deformation [i.e., $f_i = 0$ for all $i>0$ in (38)]. A Lie superalgebra is defined to be rigid, if all its deformations are trivial.

Proposition 1 (Ref. 7): If $H^2(L,L)=\{0\}$, then L is rigid.

Proof: Let $\{f_\lambda\}$ be a one-parameter family of deformations such that,

$$f_\lambda(a,b) = [a,b] + \lambda^n f_n(a,b) + \cdots,$$

i.e. $f_n (n \geqslant 1)$ is the first nonzero term in this development. Then,

$$df_n = \sum_{\text{cycl.}} \sum_{\substack{i+j=n \\ i,j<n}} (-1)^{|a||c|} f_i(f_j(a,b),c) = 0.$$

Hence, f_n is a two-cocycle. Since $H^2(L,L)=\{0\}$, there is a one-cochain g such that $f_n = dg$. Let $\phi(\lambda) = 1 + \lambda g$. Then it is verified that

$$\phi_\lambda^{-1}(f_\lambda(\phi(\lambda)x,\phi(\lambda)\cdot y)) = [x,y] + \lambda^{n+1} f'_{n+1}(x,y) + \cdots$$

where $x,y \in V$ and $f'_{n+1}, f'_{n+2}, \ldots$, are the two-cochains defining a deformation $\{f'_\lambda\}$ that is equivalent to $\{f_\lambda\}$. Since we can show that f'_{n+1} (like f_n) is a two-cocycle we prove the proposition by induction.

Let us now briefly discuss a geometric interpretation of deformations. We choose some fixed basis on V. Then, any Lie superalgebra structure on V (or V_k) is characterized by a set of structure constants with respect to a fixed basis. The graded Jacobi identities are equivalent to certain polynomial equations in several variables. Following Gerstenhaber,[7,8] one considers the algebraic manifold T defined by the solutions of these polynomial equations. Each point on T represents a Lie superalgebra, given by the structure constants obtained as solutions of the aforementioned equations, with a fixed basis. A one parameter deformation, given by (38), is an analytic curve ($\lambda \to f_\lambda$ being a polynomial function) on T.

Contraction: Let $\psi_\lambda = u + \lambda v$, where u and v are even linear mappings,[16] $V \to V$ such that ψ_λ is invertible if $\lambda \neq 0$, and for $\lambda = 0$, $\psi_0 = u$ is singular. Let us further assume that for all $x,y \in V$ the limit

$$[x,y]' = \lim_{\lambda \to 0} \psi_\lambda^{-1}[\psi_\lambda(x),\psi_\lambda(y)] \tag{46}$$

exists. Then $(x,y) \to [x,y]'$ defines a new superalgebra structure on V. Let us denote $L' = \{V,[\,,\,]'\}$. Here, L' is said to be a contraction of L. The conditions for the existence of the limit in (46) are known for the case of ordinary Lie algebra.[9]

Let us now suppose that K is a subspace of V and S is a subspace complementary to K. Let $\psi_\lambda = I_K + \lambda I_S$, where I_K (resp. I_S) is the projection operator onto K (resp. S). It is a straightforward matter to show that the limit in (46) exists if and only if K is a subalgebra of L. The resulting contraction is called Inönu–Wigner (IW)[9] contraction. This definition of IW contraction is identical to that for ordinary Lie algebras. We define a generalized IW contraction for Lie superalgebras as follows.

Let L, K, and S be as above. We have

$$L = K_{\bar 0} \oplus K_{\bar 1} \oplus S_{\bar 0} \oplus S_{\bar 1}, \tag{47}$$

where the rhs is a direct sum (internal) of subspaces. Let $I_{K_{\bar 0}}$, $I_{K_{\bar 1}}$, $I_{S_{\bar 0}}$, and $I_{S_{\bar 1}}$ denote the projection operators onto the respective subspaces in (47). Let

$$\psi_\lambda = I_{K_{\bar 0}} + I_{K_{\bar 1}} + \lambda I_{S_{\bar 1}} + \lambda^2 I_{S_{\bar 0}}, \tag{48}$$

$$[x,y]' = \lim_{\lambda \to 0}[x,y]_\lambda = \lim_{\lambda \to 0} \psi_\lambda^{-1}[\psi_\lambda(x),\psi_\lambda(y)]. \tag{49}$$

The limit in (49) exists since on expansion we get

$$[x,y]_\lambda = f_0(x,y) + \lambda f_1(x,y) + \lambda^2 f_2(x,y) + \lambda^3 f_3(x,y) + \lambda^4 f_4(x,y), \tag{50}$$

where f_i's $(i=0,1,2,3,4)$ are bilinear maps $V \times V \to V$ and $f_0(x,y) = [x,y]'$ defines the new Lie superalgebra structure on V, Comparing (50) and (38) we see that L is a one-parameter deformation of L'. Moreover, from (49), we see that for $\lambda \neq 0$ the deformation is trivial. Therefore, to find out all the Lie superalgebras that can be contracted to a given Lie algebra L', we look for the possible one-parameter deformations of L'.

VI. ILLUSTRATION

We apply the results of the previous sections to two Lie superalgebras which are of physical interest.

(a) Let $B = \mathrm{osp}(1,2)$ [or graded su(2)] spanned by $J_\pm, J_3 \in B_{\bar 0}$ and $V_{\pm 1/2} \in B_{\bar 1}$ (Ref. 17). The commutation relations are

$$\begin{aligned} &[J_3, V_{\pm 1/2}] = \pm V_{\pm 1/2}, \\ &[J_+, V_{1/2}] = [J_-, V_{-1/2}] = 0, \\ &[J_+, V_{-1/2}] = V_{1/2}, \quad \text{and} \quad [J_-, V_{1/2}] = V_{-1/2}. \end{aligned} \tag{51}$$

The even part, $B_{\bar 0} = \mathrm{su}(2)$. We also have

$$[V_{1/2}, V_{-1/2}] = \tfrac{1}{2} J_+, \quad [V_{-1/2}, V_{-1/2}] = -\tfrac{1}{2} J_-$$

and

$$[V_{1/2}, V_{-1/2}] = -\tfrac{1}{2} J_3. \tag{52}$$

Referring to (46) we let $u = I_{B_{\bar 0}}$ and $v = I_{B_{\bar 1}}$ be the respective projection operators. The contraction is of IW type. Let the contracted algebra be A. Instead of introducing a new bracket, we replace $V_{1/2}$ and $V_{-1/2}$ by α and β, respectively,

and use the same bracket, [,] to denote the Lie superalgebra product in A and B. Thus, as elements of the underlying vector space, $V_{\pm 1/2}$ are identical to α and β, respectively, but the product relations are different. The equations (51) remain intact in A (with α and β written in place of $V_{\pm 1/2}$) but Eq. (52) reduces to

$$[\alpha,\alpha] = [\beta,\beta] = [\alpha,\beta] = 0. \tag{53}$$

Therefore, $A = \{J_{\pm}, J_3, \alpha, \beta\}$ is the "full BRS algebra." This is the algebra of infinitesimal generators of global transformations that leave the Yang–Mills' Lagrangian in the Landau gauge invariant.[18] The subalgebra of A, generated by $\{J_3, \alpha, \beta\}$ is isomorphic to the BRS algebra.

Now, we address ourselves to the following question. What are the algebras that can be contracted to the full BRS algebra A? For this purpose, we look for possible deformations of A. Consequently, we have to compute $H^2(A,A)_0$, since the first coefficient f_1 in the deformation series (38) must be a two-cocycle, determined up to a coboundary. We also compute $H^3(A,A)$ to find out the obstructions to the deformation.

Computation of $H^2(A,A)$: The odd subspace $A_{\bar{1}}$ of A is an ideal of A. Moreover, $A/A_{\bar{1}} \simeq A_{\bar{0}}$ is a simple Lie algebra and hence strongly semisimple, *a fortiori.* Therefore, Theorem 6 is applicable. We have,

$$H^2(A,A) \simeq \sum_{\substack{i+j=2 \\ i,j>0}} H^i(A_{\bar{0}},\mathcal{F}) \times (H^j(A_{\bar{1}},A))^A. \tag{54}$$

In (54) we have $H^1(A_{\bar{0}},\mathcal{F}) = H^2(A_{\bar{0}},\mathcal{F}) = \{0\}$, since $A_{\bar{0}}$ is strongly semisimple. Moreover, we have to compute only the even part[16] of $H^2(A,A)$. Thus

$$\begin{aligned} H^2(A,A) &\simeq H^0(A_{\bar{0}},\mathcal{F}) \times (H^2(A_{\bar{1}},A_{\bar{0}}))^A \\ &= \mathcal{F} \times (H^2(A_{\bar{1}},A_{\bar{0}}))^A \\ &\simeq (H^2(A_{\bar{1}},A_{\bar{0}}))^A. \end{aligned} \tag{55}$$

Next, we note that $B^2(A_{\bar{1}},A_{\bar{0}}) = \{0\}$ since $A_{\bar{1}}$ is Abelian and hence,

$$H^2(A,A)_{\bar{0}} = (Z^2(A_{\bar{1}},A_{\bar{0}}))^A. \tag{56}$$

Let $f \in (Z^2(A_{\bar{1}},A_{\bar{0}}))^A$, i.e., $f: A_{\bar{1}} \times A_{\bar{1}} \to A_{\bar{0}}$ is a two-cocycle. In general, f has the form

$$\begin{aligned} f(\alpha,\alpha) &= a_1 J_+ + a_2 J_- + a_3 J_3, \\ f(\beta,\beta) &= b_1 J_+ + b_2 J_- + b_3 J_3, \end{aligned}$$

and

$$f(\alpha,\beta) = c_1 J_+ + c_2 J_- + c_3 J_3, \tag{57}$$

where $a,b,c \in \mathcal{F}$, the base field.

The cocycle condition implies

$$\begin{aligned} df(\alpha,\alpha,\alpha) &= 3\alpha \cdot f(\alpha \cdot \alpha) = 0 \\ df(\beta,\beta,\beta) &= 3\beta \cdot f(\beta,\beta) = 0, \\ df(\alpha,\alpha,\beta) &= 2\alpha \cdot f(\alpha,\beta) + \beta \cdot f(\alpha,\alpha) = 0, \end{aligned}$$

and

$$df(\alpha,\beta,\beta) = 2\beta \cdot f(\alpha,\beta) + \alpha \cdot f(\beta,\beta) = 0. \tag{58}$$

Substituting (57) for F in (58), we find that the only nonzero coefficients are a_1, b_2, and c_3. Moreover, $b_2 = c_3 = -a_1$. Hence, f has the following form:

$$\begin{aligned} &f(\alpha,\alpha) = aJ_+, \quad f(\beta,\beta) = -aJ_-, \\ &\text{and } f(\alpha,\beta) = -aJ_3, \end{aligned} \tag{59}$$

where a is a arbitrary scalar. We can also check that f is annihilated by A, i.e., $x \cdot f$ is a coboundary for all $x \in A$. Let us extend f to $A \times A$, by requiring that f be given (59) for the arguments belonging to $A_{\bar{1}} \times A_{\bar{1}}$ and f is zero otherwise.

Computation of $H^3(A,A)$: Using Theorem 6 and the fact that $(H^0(A_{\bar{1}},A))^A = 0$ we get

$$\begin{aligned} H^3(A,A)_{\bar{0}} &\simeq H^0(A_0,) \times (H^3(A_{\bar{1}},A)_{\bar{0}})^A \\ &\simeq (H^3(A_{\bar{1}},A_{\bar{1}}))^A. \end{aligned} \tag{60}$$

Now, $C^2(A_{\bar{1}},A)_{\bar{0}}$ is the space of symmetric bilinear maps $A_{\bar{1}} \times A_{\bar{1}} \to A_{\bar{0}}$ and $\dim C^2(A_{\bar{1}},A_{\bar{0}}) = \dim (\mathrm{Sym.}(A_{\bar{1}} \otimes A_{\bar{1}})) \times \dim A_{\bar{0}} = 9$. Similarly, $C^3(A_{\bar{1}},A_{\bar{1}})$ is the space of symmetric functions $A_{\bar{1}} \times A_{\bar{1}} \times A_{\bar{1}} \to A_{\bar{1}}$. This space has dimension = 8. The coboundary operator d maps $C^2(A_{\bar{1}},A_{\bar{0}}) \to C^3(A_{\bar{1}},A_{\bar{1}})$ and $\dim(\mathrm{Ker}\, d) = \dim Z^2(A_{\bar{1}},A_{\bar{0}}) = 1$, by the preceding computation [cf. (59)]. This implies that $H^3(A_{\bar{1}},A_{\bar{1}}) = \{0\}$. Hence, every two-cocycle is integrable. But, the only even[16] two-cocycle (up to a scalar multiple) is given by (59). Therefore, putting $a = \frac{1}{2}$ in (59) we denote the corresponding function by f_1. We have, $f_1: A_{\bar{0}} \times A_{\bar{0}} \in 0$ and $f_1: A_{\bar{0}} \times A_{\bar{1}} \to 0$ and restriction of f_1 to $A_{\bar{1}} \times A_{\bar{1}}$ is given by (59) with $a = \frac{1}{2}$. We see that $f_1(f_1(x,y),z) = 0$ for all $x,y,z \in A$. Hence, we can assume that $f_2 = f_3 = \cdots = 0$ in (38). We have a first-order deformation with

$$f_\lambda = f_0 + \lambda f_{\bar{1}}, \tag{61}$$

where f_0 is the original multiplication on A and we get an algebra that is isomorphic to osp(1,2). As osp(1,2) is known to be strongly semisimple[19] [$\because$ osp(1,2) = $B(0,1)$ no further nontrivial deformation is possible (cf. Proposition 1)]. We therefore, conclude that osp(1,2) is the only algebra that can be contracted to the full BRS algebra.

(b) The super Poincaré algebra. Let L denote the super Poincaré algebra. We wish to determine all the nonisomorphic Lie superalgebras that can be contracted to L. Thus we have to determine the possible deformations of L. Let V be the underlying vector space and $\mathcal{F}$ the base field of L.

Here, $L_{\bar{0}}$ is the Poincaré algebra. Let us write $L_{\bar{0}} = D +)P$, where $D = R^1 \oplus R^2$ with $R^1 \simeq R^2 \simeq \mathrm{su}(2)$, $+$) denotes the semidirect sum and $P = \{p_1,p_2,p_3,p_4\}$, is the ideal generated by infinitesimal translation operators. Also, D is the Lorentz algebra. Let $\{h_i,x_i,y_i | i = 1,2\}$ be the basis for R^i with the commutation relations given by $[h_1,x_1] = 2x_1, [h_1,y_1] = -2y_1, [x_1,y_1] = h_1$, etc. A acts on P in the adjoint representation with highest weight μ, such that $\mu(h_i) = 1$; $L_{\bar{1}} = \{\alpha_i,\beta_i | i = 1,2\}$. The α_i's (resp. β_i) span an R^1 module (resp. R^2 module) in the adjoint representation: e.g., $[h_1,\alpha_1] = \alpha_1$, $[h_1,\alpha_2] = \alpha_2$, etc., and $[R^1,\beta_i] = [R^2,\alpha_i] = 0$, $i = 1,2$ and $[P,L_{\bar{1}}] = 0$. Finally, we have

$$\begin{aligned} &[\alpha_1,\beta_1] = p_1, \quad [\alpha_1,\beta_2] = p_2, \\ &[\alpha_2,\beta_1] = p_3, \end{aligned}$$

and

$$[\alpha_2,\beta_2] = p_4. \tag{62}$$

The basis used for the super Poincaré algebra L, differs from the one found in the literature. We have avoided the 'dot' notation.[20] We are also reminded that, in (62) [,] is the anticommutator since we have used this bracket for the graded computator. $B = P \oplus L_{\bar{1}}$ (vector space direct sum) is an ideal in L. The quotient algebra $L/B \simeq D$ is strongly semisimple ($\because$ it is a semisimple Lie algebra). Therefore, Theorem 6 can be used to compute the cohomology groups. We will omit the lengthy details of calculations that will appear elsewhere.[21] The final results are mentioned below.

Now, $H^2(L,L)_{\bar{0}} \simeq \mathcal{F} \times \mathcal{F}$ and a typical representative cocycle, f_1 is given by

$$f_1(\alpha_1,\alpha_1) = 2bx_1, \quad f_1(\alpha_2,\alpha_2) = -2by_1,$$
$$f_1(\alpha_1,\alpha_2) = bh_1,$$
$$f_1(\beta_1,\beta_1) = 2cx_2, \quad f_1(\beta_2,\beta_2) = -2cy_2,$$

and

$$f_1(\beta_1,\beta_2) = -ch_2, \tag{63}$$

where $a,b,c,\in\mathcal{F}$. Further,

$$\begin{aligned} f_1(p_3,\alpha_1) &= -b\beta_1, & f_1(p_4,\alpha_1) &= -b\beta_2, \\ f_1(p_1,\alpha_2) &= -b\beta_1, & f_1(p_1,\alpha_2) &= b\beta_2, \\ f_1(p_2,\beta_1) &= -c\alpha_1, & f_1(p_4,\beta_1) &= -c\alpha_2, \\ f_1(p_1,\beta_2) &= c\alpha_1, & f_1(p_3,\beta_1) &= c\alpha_2, \end{aligned} \tag{64}$$

and f_1 is zero on all other elements of $L \times L$. Now, let us recall that R^1 and R^2 are not independent but conjugates of each other (dotted and undotted spinors!) and let $\mathcal{F} = R$, the real field. Then, $c = \bar{b} = b$ in (63) and (64).

The computation of $H^3(L,L)$ is more involved but, it can be shown to be nonzero. However, f_1 is integrable and the integrability conditions determine f_2, the second-order coefficient in the deformation series (38). Thus,

$$f_2(p_1,p_2) = ax_1, \quad f_2(p_1,p_3) = ax_2,$$
$$f_2(p_1,p_4) = -(a/2)(h_1 + h_2),$$
$$f_2(p_2,p_3) = (a/2)(h_1 - h_3),$$
$$f_2(p_2,p_4) = -ay_2,$$

and

$$f_2(p_3,p_4) = -ay_1, \tag{65}$$

where $a = b^2$.

Hence, we have a second-order deformation of the super Poincaré algebra,

$$f_\lambda(x,y) = [x,y] + \lambda f_1[x,y] + \lambda^2 f_2(x,y), \tag{66}$$

where $x,y \in V \cdot f_\lambda$ defines a new Lie superalgebra structure on V.

The scalar b in (63) and (64) can be taken as the deformation parameter. Then, from (66) (with $\lambda = b$) we obtain the commutation relation for the new Lie superalgebra T on the vector space V. From Eqs. (61)–(66) one identifies T with the super de Sitter algebra and b^2 is now identified with scalar curvature of a de Sitter universe. Moreover, we can also conclude that the super de Sitter algebra is the only Lie superalgebra (up to an isomorphism) that can be contracted to the super Poincaré algebra. The corresponding contraction is the generalized IW contraction discussed at the end of the previous section. Unlike the nonsupersymmetric case,[8] where both so(4,1) and so(3,2) can be contracted to yield the Poincaré algebra, we have here a unique algebra T such that $T_{\bar{0}} \simeq$ so(4,1) only. The curvature b^2 is positive and hence, rules out negative gravity!

ACKNOWLEDGMENTS

One of the authors (K.C.T) would like to thank Professor Abdus Salam, the International Atomic Energy Agency, and UNESCO for hospitality at the International Centre for Theoretical Physics, Trieste. He would also like to thank Professor M. S. Narasimhan and Professor J. Strathdee for fruitful discussions. The valuable suggestions made by the referee, which helped us to bring the manuscript to the present form, are also gratefully acknowledged.

[1] D. Leites, Functional Anal. Appl. **9**, 75 (1975).
[2] B. R. Sitaram and K. C. Tripathy, Indian J. Pure Appl. Math. **13**, 672 (1982).
[3] B. R. Sitaram and K. C. Tripathy, J. Math. Phys. **25**, 443 (1984).
[4] B. Mitra and K. C. Tripathy, Indian J. Pure Appl. Math. **190**, 333 (1988).
[5] G. Hochschild and J. P. Serre, Ann. Math. **57**, 603 (1953).
[6] F. Bayen, M. Flato, C. Fronsdal, A. Lichnerowicz, and D. Sternheimer, Ann. Phys. **111**, 61, 151 (1978); A. Lichnerowicz, in *Quantum Theory, Fields and Particles,* Mathematical and Physics Studies, Vol. 4, edited by A. O. Barut (Reidel, Dordrecht, 1982), D. Arnal, M. Cahen, and S. Gutt, J. Geom. Phys. **3**, 327 (1986).
[7] M. Gerstenhaber, Ann. Math. **79**, 59 (1964); M. Gerstenhaber and S. D. Schack, in *Deformation Theory of Algebras and Structures and Applications,* edited by M. Hazewinkel and M. Gerstenhaber, NATO Advanced Sciences Institute Series C, Mathematical and Physical Sciences, Vol. 247 (Kluwer, Dordrecht, 1988).
[8] M. Levy-Nahas, J. Math. Phys. **8**, 1222 (1967).
[9] E. Inonu and E. P. Wigner, Proc. Natl. Acad. Sci. (USA) **39**, 510 (1953); E. J. Saletan, J. Math. Phys. **2**, 1 (1961).
[10] H. Cartan and S. Eilenberg, *Homological Algebra* (Princeton U. P., Princeton, NJ, 1956) S. Maclane, *Homology* (Springer-Verlag, Berlin, 1963).
[11] N. Jacobson, *Lie Algebras* (Interscience, New York, 1962). S. Vardrajan, *Lie Groups, Lie Algebras and their Representations* (Prentice-Hall, Englewood Cliffs, NJ, 1974).
[12] V. G. Kac, Adv. Math. **26**, 8 (1977).
[13] C. Chevalley and S. Eilenberg, Trans. Am. Math. Soc. **63**, 85 (1948).
[14] G. Hochschild, Illinois J. Math. **20**, 107 (1976); D. Z. Djokovic and G. Hochschild, *ibid.* **20**, 134 (1976).
[15] J-L. Koszul, Bull, Soc. Math. France **78**, 65 (1950); W. Greub, S. Halperin, and R. Vanstone, *Connections, Curvature and Cohomology* (Academic, New York, 1976), Vol. III, Chap. V.
[16] The Lie superalgebra produce on a Z_2-graded vector space $V = V_{\bar{0}} \oplus V_{\bar{1}}$ must be even, i.e., the product maps $V_{\bar{0}} \times V_{\bar{0}} \to V_{\bar{0}}, V_{\bar{1}} \times V_{\bar{1}} \to V_{\bar{0}}$ and $V_{\bar{0}} \times V_{\bar{1}} \to V_{\bar{1}}$. Therefore, we need consider only even bilinear maps $V \times V \to V$.
[17] M. K. Patra and K. C. Tripathy, Lett. Math. Phys. **17**, 1 (1989).
[18] K. Nishijima, Prog. Theor. Phys. **72**, 1214 (1984).
[19] A. Pais and V. Rittenberg, J. Math. Phys. **16**, 2062 (1975).
[20] S. J. Gates, Jr., M. T. Grisaru, M. Roček, and W. Siegel, *Superspace* (Benjamin/Cummings, New York, 1983).
[21] H. D. Doebner, K. C. Tripathy, and M. K. Patra, "Deformation of super Poincaré Algebra," to be submitted.

Note on asymptotic series expansions for the derivative of the Hurwitz zeta function and related functions

S. Rudaz
School of Physics and Astronomy, University of Minnesota, Minneapolis, Minnesota 55455

(Received 18 July 1989; accepted for publication 8 August 1990)

Asymptotic series for the Hurwitz zeta function, its derivative, and related functions (including the Riemann zeta function of odd integer argument) are derived as an illustration of a simple, direct method of broad applicability, inspired by the calculus of finite differences.

I. INTRODUCTION

The Hurwitz zeta function is defined for $a \neq 0, -1, -2, \ldots$, and $Re\, s > 1$ by the series

$$\zeta(s,a) = \sum_{n=0}^{\infty} \frac{1}{(n+a)^s}, \tag{1}$$

for which it is readily seen that

$$\zeta(s,a+1) - \zeta(s,a) = -a^{-s}, \tag{2}$$

$$\frac{\partial}{\partial a}\zeta(s,a) = -s\zeta(s+1,a). \tag{3}$$

When $a = 1$, the Hurwitz zeta function reduces to the Riemann zeta function, $\zeta(s,1) = \zeta(s)$. The values of $\zeta(s,a)$ and its derivative $(\partial/\partial s)\zeta(s,a) \equiv \zeta'(s,a)$ for $s = -n$, with integer $n \geqslant 0$ often arise in calculations of quantum effects in field and string theories through the method of zeta-function regularization.[1-3] Some time ago, Elizalde[4] derived an asymptotic expansion for $\zeta'(-n,a)$ when a is large that is extremely useful in the computation of effective actions in nontrivial backgrounds: The method used, similar to the conventional one that follows from Watson's lemma using Laplace's method, starts with Hermite's integral representation for $\zeta(s,a)$ (Ref. 5, p. 26) and is quite laborious. The purpose of this paper is to describe a simpler derivation of this result, using a method that is quite direct and also allows for the recovery of a number of classical results, as well as being applicable to a variety of other problems.

II. METHOD

The basic formula is obtained through a very simple observation, inspired by standard methods of the calculus of finite differences (see Ref. 6 for a comprehensive treatment): Let $f(t)$ be a sufficiently differentiable function and consider the integral

$$\int_0^1 dt\, f'(t) = f(1) - f(0). \tag{4}$$

Recall now that the Bernoulli polynomials $B_n(t)$ (Ref. 5, p. 36 and Ref. 6, pp. 17–23) are uniquely determined by the identity, for integer $n > 0$,

$$\int_x^{x+1} dt\, B_n(t) = x^n,$$

from which the all-important property

$$B_n'(t) = nB_{n-1}(t) \tag{5}$$

follows, as well as the relation $B_n(x+1) - B_n(x) = nx^{n-1}$. One readily sees that $B_0(t) = 1$, $B_1(t) = t - \frac{1}{2}$, $B_2(t) = t^2 - t + \frac{1}{6}$, $B_3(t) = t^3 - \frac{3}{2}t^2 + \frac{1}{2}t$, etc., and that for $n \geqslant 2$, $B_n(1) = B_n(0) \equiv B_n$ with B_n the well-known Bernoulli numbers. It is also easy to see, directly from their definition, that $B_n(1-t) = (-1)^n B_n(t)$ whence $B_{2n+1} = 0$ for $n > 0$.

Returning now to Eq. (4), write using Eq. (5),

$$\begin{aligned}\int_0^1 dt\, f'(t) &= \int_0^1 dt\, f'(t) B_1'(t) \\ &= \frac{1}{2}(f'(1) + f'(0)) \\ &\quad - \frac{1}{2}\int_0^1 dt\, f''(t) B_2'(t).\end{aligned} \tag{6}$$

Upon integrating by parts, continuing this process indefinitely results, after a simple rearrangement, in

$$f'(0) = \Delta(0) - \frac{1}{2}\Delta'(0) + \sum_{k=2}^{\infty} \frac{(-1)^k}{k!} B_k \Delta^{(k)}(0), \tag{7}$$

where $\Delta(t) = f(t+1) - f(t)$. Note that in this equation the infinite series is to be more properly understood as a sum over k terminating at some arbitrarily chosen finite N, together with an integral formula for the remainder R_{N+1}.

To apply this formula to the Hurwitz zeta function and its derivative, we next show that the properties (2) and (3) can be extended to the entire complex s plane. The Hurwitz zeta function can be extended as a meromorphic function to the entire complex s plane by means of the contour integral (Ref. 5, p. 25)

$$\zeta(s,a) = -\frac{\Gamma(1-s)}{2\pi i}\int_C \frac{(-z)^{s-1}e^{-az}}{1-e^{-z}}\, dz, \tag{8}$$

where the contour C forms a loop around the positive real axis. Using this representation one has

$$\begin{aligned}\zeta(s,a+1) - \zeta(s,a) &= \frac{\Gamma(1-s)}{2\pi i}\int_C (-z)^{s-1}e^{-az}\, dz \\ &= -a^{-s},\end{aligned} \tag{9}$$

upon using the reflection formula for the gamma function

 0022-2488/90/122832-03$03.00

$\Gamma(s)\Gamma(1-s) = \pi \csc \pi s$, because with the above choice of contour,

$$-\int_C (-z)^{s-1} e^{-z}\, dz = 2i \sin \pi s\ \Gamma(s)$$

is just Hankel's representation for the gamma function. As to the other relation, note that

$$\frac{\partial}{\partial a}\zeta(s,a) = -\frac{\Gamma(1-s)}{2\pi i}\int_C \frac{(-z)^s e^{-az}}{1-e^{-z}}\, dz$$
$$= -s\zeta(s+1,a), \tag{10}$$

by virtue of the functional relation for the gamma function $\Gamma(1+s) = s\Gamma(s)$.

With these results in hand, and as a first application of (7), let us quickly derive an asymptotic expansion for $\zeta(s,a)$, valid for $|\arg a| < \pi$. Put $f(t) = \zeta(s,a+t)$ and then from (9) and (10) one has immediately $f'(0) = -s\zeta(s+1,a)$, $\Delta(0) = -a^{-s}$, and $\Delta^{(k)}(0) = (-1)^{k+1}(s)_k a^{-s-k}$, where $(s)_k = s(s+1)\cdots(s+k-1) = \Gamma(s+k)/\Gamma(s)$ is Pochhammer's symbol (rising factorial function). Now one has, directly from (7), the asymptotic expansion (Ref. 5, p. 48)

$$\zeta(s+1,a) = \frac{1}{s}a^{-s} + \frac{1}{2}a^{-s-1}$$
$$+\frac{1}{s}\sum_{k=2}^{\infty}\frac{B_k}{k!}(s)_k a^{-s-k}. \tag{11}$$

This derivation has been fairly easy, and in any event is simpler and more direct than that suggested in Erdélyi's compendium.[5] When $s = -n$ for integer $n > 0$, the series (11) terminates and with $(-n)_k = (-1)^k n!/(n-k)!$ one finds

$$n\zeta(1-n,a) = -a^n + \frac{1}{2}na^{n-1}$$
$$-\sum_{k=2}^{n}(-1)^k\binom{n}{k}B_k a^{n-k}. \tag{12}$$

Recalling that $B_1 = -1/2$ and that for otherwise odd k, $B_k = 0$, this reduces to the well-known result (Ref. 5, p. 27)

$$\zeta(1-n,a) = -\frac{1}{n}\sum_{k=0}^{n}\binom{n}{k}B_k a^{n-k} = -\frac{1}{n}B_n(a) \tag{13}$$

in terms of Bernoulli polynomials.

We now turn, after these preliminaries, to the problem at hand, namely, the derivation of the asymptotic expansion for $\zeta'(s,a) \equiv (\partial/\partial s)\zeta(s,a)$ as it follows using again formula (7). Note that while one could consider simply differentiating Eq. (11) term by term, such a procedure is not generally valid and would have to be justified. Instead, we proceed directly, putting $f(t) = \zeta'(s,a+t)$ whence it follows that $f'(t) = -s\zeta'(s+1,a+t) - \zeta(s+1,a+t)$ and $\Delta(t) = (a+t)^{-s}\log a + t$. After a simple calculation using Leibniz' rule and Eq. (11), one finds by Eq. (7)

$$\zeta'(s+1,a) = -\zeta(s+1,a)\log a$$
$$-(1/s)\left[\zeta(s+1,a) - \tfrac{1}{2}a^{-s-1}\right]$$
$$+\frac{1}{s}\sum_{k=2}^{\infty}B_k\sum_{j=0}^{k-1}\frac{(s)_j}{j!(k-j)}a^{-s-k}. \tag{14}$$

It is straightforward to check that this is also the result that follows from term-by-term differentiation of (11), a procedure which is thus justified *a posteriori*.

Now set $s = -n$ for integer $n > 0$ in (14). This becomes, with $(-n)_j = (-1)^j\Gamma(n+1)/\Gamma(n+1-j)$,

$$\zeta'(1-n,a) = -\zeta(1-n,a)\log a$$
$$+(1/n)\left[\zeta(1-n,a) - \tfrac{1}{2}a^{n-1}\right]$$
$$-\frac{1}{n}\sum_{k=2}^{\infty}B_k\sum_{j=0}^{M}\binom{n}{j}\frac{(-1)^j}{k-j}a^{n-k}, \tag{15}$$

with $M = \min\ (n, k-1)$. When $n = 1$, $M = \min\ (1, k-1) = 1$ for all k and one can use the identity

$$\sum_{j=0}^{n}\binom{n}{j}\frac{(-1)^j}{k-j} = \frac{(-1)^n n!}{k(k-1)\cdots(k-n)} \tag{16}$$

to directly obtain, using (13),

$$\zeta'(0,a) = a\log a - a - \frac{1}{2}\log a + \sum_{k=2}^{\infty}\frac{B_k}{k(k-1)}a^{1-k}. \tag{17}$$

When $n > 1$, the double sum in (15) can be reexpressed as

$$\sum_{k=2}^{\infty}\sum_{j=0}^{M} = \sum_{k=2}^{n}\sum_{j=0}^{k-1} + \sum_{k=n+1}^{\infty}\sum_{j=0}^{n}$$

and again using (13) and (16) one has the general formula, for $n > 1$,

$$\zeta'(1-n,a) = \frac{1}{n}B_n(a)\log a - \frac{1}{n^2}\left[B_n(a) + \frac{n}{2}a^{n-1}\right]$$
$$-\frac{1}{n}\sum_{k=2}^{n}B_k\sum_{j=0}^{k-1}\binom{n}{j}\frac{(-1)^j}{k-j}a^{n-k}$$
$$+(-1)^{n-1}(n-1)!\sum_{k=n+1}^{\infty}$$
$$\times\frac{B_k}{k(k-1)\cdots(k-n)}a^{n-k}. \tag{18}$$

Explicitly, one has for example for $n = 2$ and 3 the expressions,

$$\zeta'(-1,a) = \frac{1}{2}\left[a^2 - a + \frac{1}{6}\right]\log a - \frac{1}{4}a^2 + \frac{1}{12}$$
$$-\sum_{k=3}^{\infty}\frac{B_k}{k(k-1)(k-2)}a^{2-k}, \tag{19}$$

$$\zeta'(-2,a) = \frac{1}{3}\left[a^3 - \frac{3a^2}{2} + \frac{a}{2}\right]\log a - \frac{1}{9}a^3 + \frac{1}{12}a$$
$$+2\sum_{k=4}^{\infty}\frac{B_k}{k(k-1)(k-2)(k-3)}a^{3-k}, \tag{20}$$

and so on. In a slightly more transparent notation, Eq. (18) is precisely Elizalde's asymptotic expansion[4] for large $|a|$, $|\arg a| < \pi$.

Quite apart from its application in the calculation of effective actions in quantum field and string theories, formula (18) can be used to obtain asymptotic series for the

Riemann zeta function of odd positive integer argument. Indeed, put $s=-2n+\epsilon$ (with integer $n>0$) in the functional equation for the Riemann zeta function $\zeta(s)=\zeta(s,1)$, namely (Ref. 5, p. 35),

$$\zeta(1-s)=2(2\pi)^{-s}\Gamma(s)\zeta(s)\cos \pi s/2.$$

Now taking the limit $\epsilon\to 0$ and noting that $\zeta(-2n)=0$, it readily follows that

$$\zeta(2n+1)=[2(2\pi)^{2n}/(2n)!](-1)^n\zeta'(-2n). \quad (21)$$

Taking $\zeta'(-2n)=\zeta'(-2n,1)$ in this equation one gets an asymptotic series for $\zeta(2n+1)$, with integer $n>0$, namely,

$$\zeta(2n+1)\sim\frac{(-1)^{n+1}(2\pi)^{2n}}{(2n+1)!}\times\left[1+2\sum_{k=2}^{2n+1}\sum_{j=0}^{k-1}\binom{2n+1}{j}\frac{(-1)^j B_k}{k-j}\right]+(-1)^n 2(2\pi)^{2n}\times\sum_{k=2n+2}^{\infty}\frac{B_k}{k(k-1)\cdots(k-2n-1)}. \quad (22)$$

For example, one gets for $\zeta(3)$,

$$\zeta(3)\sim\frac{\pi^2}{9}-8\pi^2\sum_{k=4}^{\infty}\frac{B_k}{k(k-1)(k-2)(k-3)}. \quad (23)$$

Here the symbol "$\sim$" has its usual meaning in this context, namely, that the first n terms of the asymptotic series on the right-hand side of Eq. (23) yield an approximation to $\zeta(3)$ with an error less than the magnitude of the $(n+1)$th term. Under these conditions, the best approximation to $\zeta(3)$ is obtained by keeping five terms in the series, with the result 1.2014. In view of the fact that the Bernoulli numbers alternate in sign, it is sensible to average this with the result of keeping only four terms in the series, which yields the mean approximation 1.2020, to be compared with the tabulated value of $\zeta(3)=1.20205690\cdots$.

III. IN LIEU OF A CONCLUSION

We conclude with some expository remarks, illustrating some further applications of Eq. (7). The logarithm of the gamma function satisfies the finite difference equation

$$\log\Gamma(a+1)-\log\Gamma(a)=\log a.$$

Now put $f'(t)=\log\Gamma(a+t)$ and then $\Delta'(t)=\log a+t$. Integrating, it follows that $\Delta(0)=a\log a-a+c$ where c is an integration constant, and one has immediately from (7)

$$\log\Gamma(a)=a\log a-a+c-\tfrac{1}{2}\log a+\sum_{k=2}^{\infty}\frac{B_k}{k(k-1)}a^{1-k}. \quad (24)$$

An additional piece of information is required to determine c: This can be done by inserting Legendre's duplication formula

$$\Gamma(a)\Gamma(a+\tfrac{1}{2})=\sqrt{\pi}2^{1-2a}\Gamma(2a)$$

into Eq. (24) and matching terms not depending on a. The result is

$$c=\tfrac{1}{2}\log 2\pi. \quad (25)$$

With this, (24) is just Stirling's asymptotic expansion for the logarithm of the gamma function (Ref. 5, p. 47). Note that differentiating (9) with respect to s and setting $s=0$ yields

$$\zeta'(0,a+1)-\zeta'(0,a)=\log a,$$

while from (9) and (10) one also finds that

$$\frac{d^2}{da^2}\zeta'(0,a)=\zeta(2,a)>0.$$

Precisely the same equations apply with $\zeta'(0,a)$ replaced by $\log\Gamma(a)$ and are the basis of Artin's proof of the uniqueness of the gamma function.[7] It thus follows that $\zeta'(0,a)$ and $\log\Gamma(a)$ can differ at most by an additive constant. In fact, by comparing the asymptotic expansions (17) and (24), and given (25), we recover the result

$$\log\Gamma(a)=\zeta'(0,a)+\tfrac{1}{2}\log 2\pi$$

a classical formula due to Lerch (Ref. 8, p. 271). Finally, setting $a=1$, one also sees that $\zeta'(0)=-\tfrac{1}{2}\log 2\pi$.

ACKNOWLEDGMENTS

This work was supported in part by the Department of Energy, under contract DE-AC-02-83ER-40105 and by a Presidential Young Investigator Award.

[1] A. Salam and J. Strathdee, Nucl. Phys. B **90**, 203 (1975).
[2] J. S. Dowker and R. Critchley, Phys. Rev. D **13**, 3224 (1976).
[3] S. W. Hawking, Commun. Math. Phys. **55**, 133 (1977).
[4] E. Elizalde, Math. Comput. **47**, 347 (1986).
[5] *Higher Transcendental Functions*, edited by A. Erdélyi (McGraw-Hill, New York, 1953), Vol. 1.
[6] N. E. Nörlund, *Vorlesungen über Differenzenrechnung* (Springer, Berlin, 1924).
[7] E. Artin, *Einführung in die Theorie der Gammafunktion* (Teubner, Leipzig, 1931).
[8] E. T. Whittaker and G. N. Watson, *A Course of Modern Analysis* (Cambridge U.P., Cambridge, 1969), 4th ed.

Integrable Hamiltonian systems related to the polynomial eigenvalue problem

Yunbo Zeng and Yishen Li
Department of Mathematics, University of Science and Technology of China, Hefei, 230026, People's Republic of China

(Received 18 January 1990; accepted for publication 20 June 1990)

The independent integrals of motion in involution for the Hamiltonian system related to the second-order polynomial eigenvalue problem are constructed by using relevant recursion formula. The hierarchy of Hamiltonian systems obtained from the above problem and the time part of the Lax pair are shown to be completely integrable and they are shown to commute with each other. Furthermore, their solution solves the evolution equation associated with the Lax pair.

I. INTRODUCTION

There are some ways to restrict infinite-dimensional integrable Hamiltonian systems to finite-dimensional invariant submanifolds of their phase space in order to obtain finite-dimensional integrable Hamiltonian systems (see, for example, Refs. 1–5). We have proposed in Refs. 6 and 7 a straightforward way to obtain a hierarchy of finite-dimensional integrable Hamiltonian systems by restricting a hierarchy of integrable evolution equations to the invariant subspace of their recursion operator. The independent integrals of motion for these Hamiltonian systems can be constructed by using the recursion formula related to the associated eigenvalue problem, and can be shown to be in involution. Thus all of these systems are completely integrable Hamiltonian systems in the sense of Liouville[8] and commute with each other. The solution to these Hamiltonian systems solves the evolution equation.

In the present paper, we consider the second-order polynomial eigenvalue problem given by (2.1). We first obtain a natural constraint on potential by restricting it on the linear space spanned by the eigenfunctions of the recursion operator in Sec. II. Under this constraint condition, the system stemming from the eigenvalue problem becomes a Hamiltonian system. In Sec. III, by using the relevant recursion formula, the integrals of the motion for this system can be constructed and shown to be in involution. Thus this Hamiltonian system is completely integrable in the sense of Liouville. Finally, in Sec. IV, the systems obtained from the time part of the Lax pair are found to be completely integrable and commute with each other.

II. THE CONSTRAINT ON POTENTIAL

Consider the second-order polynomial eigenvalue problem[9,10]

$$\phi_{xx} + \sum_{i=0}^{m-1} \lambda^i u_i \phi = \lambda^m \phi, \tag{2.1}$$

in which isospectral flows are shown to possess $(m+1)$ compatible Hamiltonian structures.[11] The time evolution equation for ϕ is taken to be

$$\phi_{t_n} = -\tfrac{1}{2} B_x^{(n)} \phi + B^{(n)} \phi_x, \tag{2.2}$$

where

$$B^{(n)} = \sum_{k=0}^{n} b_k \lambda^{n-k}, \quad b_0 = 1,$$

$$R_k = (R_k^{(0)},...,R_k^{(m-2)}, 2b_{k+1})^T, \quad u = (u_0,...,u_{m-1})^T, \tag{2.3a}$$

$$R_k = LR_{k-1} = L^k u,$$

$$L = \begin{pmatrix} 0 \cdots 0 & J_0 \\ 1 \cdots 0 & J_1 \\ 0 \cdots 1 & J_{m-1} \end{pmatrix}, \tag{2.3b}$$

$$J_0 = \tfrac{1}{4} D^2 + u_0 - \tfrac{1}{2} D^{-1} u_{0x},$$

$$J_i = u_i - \tfrac{1}{2} D^{-1} u_{ix}, \quad i = 1,...,m-1.$$

Here, $D = \partial/\partial x$, the integral constant of the integral operator D^{-1} appearing in L is defined to be zero. Then the hierarchy of evolution equations,[10,11] deduced from the solvability condition of (2.1) and (2.2), can be rewritten as

$$u_{t_n} = DL^n u. \tag{2.4}$$

Also, from the solvability condition, it is found[11] that b_k satisfies

$$b_k = \sum_{i=0}^{m-1} J_i b_{k-m+i}, \tag{2.5}$$

and if ϕ is a solution of (2.1), then

$$D \sum_{i=0}^{m-1} \lambda^i J_i \phi^2 = \lambda^m (\phi^2)_x. \tag{2.6}$$

By using (2.6) another recursion formula for b_k without the integral operator D^{-1} can be found. Rewrite (2.6) as

$$D \sum_{i=0}^{m-1} \lambda^i J_i P = \lambda^m P_x. \tag{2.7}$$

Inserting the expansion

$$P = \sum_{j=0}^{\infty} P_j \lambda^{-j} \tag{2.8}$$

into (2.7), we find that P_j satisfy same relationship (2.5) as b_j do. Multiplying both sides of (2.7) by P and integrating it once yield

$$P_{xx} P - \frac{1}{2} (P_x)^2 + 2 \sum_{i=0}^{m-1} \lambda^i u_i P^2 = 2\lambda^m P^2 + C. \tag{2.9}$$

Set $P_0 = 1$, $C = -2\lambda^m$. By substituting (2.8) into (2.9), we find that $P_j = b_j$,

 0022-2488/90/122835-05$03.00

$$b_k=\frac{1}{2}\sum_{i=1}^{k}u_{m-i}\sum_{j=0}^{k-i}b_jb_{k-i-j}-\frac{1}{2}\sum_{j=1}^{k-1}b_jb_{k-j},\quad k\leqslant m, \tag{2.10a}$$

$$\begin{aligned}b_{k+m}=&\frac{1}{4}\sum_{j=1}^{k}b_{jxx}b_{k-j}-\frac{1}{8}\sum_{j=1}^{k-1}b_{jx}b_{k-j,x}\\&+\frac{1}{2}\sum_{i=0}^{m-1}u_i\sum_{j=0}^{k+i}b_jb_{k+i-j}\\&-\frac{1}{2}\sum_{j=1}^{k+m-1}b_jb_{k+m-j},\quad k=1,2,\ldots.\end{aligned} \tag{2.10b}$$

We now consider the following system instead of (2.1):

$$\phi_{jxx}+\sum_{i=0}^{m-1}\lambda_j^iu_i\phi_j=\lambda_j^m\phi_j,\quad j=1,\ldots,N, \tag{2.11}$$

where $\lambda_j\neq\lambda_k$ when $j\neq k$. It is found from (2.6) that if $q=(q_1,\ldots q_N)^T=(\phi_1,\ldots,\phi_N)^T$ is a solution of (2.11), then we have

$$DL\Psi_j=\lambda_j\Psi_{jx},\quad j=1,\ldots,N, \tag{2.12a}$$

where

$$\begin{aligned}&\Psi_j=(\Psi_j^{(0)},\Psi_j^{(1)},\ldots,\Psi_j^{(m-1)})^T,\\&\Psi_j^{(0)}=\lambda_j^{m-1}\phi_j^2-\sum_{i=1}^{m-1}\lambda_j^{i-1}J_i\phi_j^2,\\&\Psi_j^{(m-k)}=\lambda_j^{k-1}\phi_j^2-\sum_{i=1}^{k-1}\lambda_j^{k-1-i}J_{m-i}\phi_j^2,\\&\Psi_j^{(m-1)}=\phi_j^2.\end{aligned} \tag{2.12b}$$

Throughout this paper no boundary condition on u and q is imposed. Thus (2.12a) leads to

$$L\Psi_j=\lambda_j\Psi_j+\sum_{i=0}^{m-1}\beta_ie_i, \tag{2.13a}$$

where β_i are some constants, $e_0=(1,0,\ldots,0)^T,\ldots,e_{m-1}=(0,\ldots,0,1)^T$. Notice that

$$L\sum_{i=0}^{m-1}\beta_ie_i=\sum_{i=0}^{m-2}\beta_ie_{i+1}+\frac{1}{2}\beta_{m-1}u, \tag{2.13b}$$

if we take

$$u=\sum_{j=1}^{N}\Psi_j, \tag{2.14a}$$

we find from (2.13) that the linear space M spanned by $\{\Psi_1,\ldots,\Psi_N,e_0,\ldots e_{m-1}\}$ is the invariant subspace of L. This property plays important role in our approach.

Proposition 1: The constraint on potential obtained from (2.14a) is of the form

$$\begin{aligned}&u_{m-k}=\sum_{j=1}^{k}a_j\sum_{l_1+\cdots+l_j=k-j}\langle\Lambda^{l_1}q,q\rangle\cdots\langle\Lambda^{l_j}q,q\rangle,\\&k=1,\ldots,m,\end{aligned} \tag{2.14b}$$

where $l_i\geqslant 0$, $\langle\cdot,\cdot\rangle$ is the inner product in $\mathbb{R}^N$, and

$$a_0=-1,$$

$$a_j=-\frac{j+1}{2j}a_{j-1}=(-1)^j\frac{j+1}{2j},\quad j=0,1,2,\ldots.$$

$$\Lambda=\operatorname{diag}(\lambda_1,\ldots,\lambda_N).$$

Proof: It is clear that (2.14b) holds for $k=1$. From (2.12b) and (2.14a), we have by induction that

$$\begin{aligned}u_{m-k}&=\langle\Lambda^{k-1}q,q\rangle-\sum_{i=1}^{k-1}u_{m-i}\langle\Lambda^{k-1-i}q,q\rangle+\frac{1}{2}D^{-1}\sum_{i=1}^{k-1}u_{m-i,x}\langle\Lambda^{k-1-i}q,q\rangle\\&=\langle\Lambda^{k-1}q,q\rangle-\sum_{i=1}^{k-1}\langle\Lambda^{k-1-i}q,q\rangle\sum_{j=1}^{i}a_j\sum_{l_1+\cdots+l_j=i-j}\langle\Lambda^{l_1}q,q\rangle\cdots\langle\Lambda^{l_j}q,q\rangle+\frac{1}{2}D^{-1}\sum_{i=1}^{k-1}\langle\Lambda^{k-1-i}q,q\rangle\\&\quad\times\sum_{j=1}^{i}a_j\sum_{l_1+\cdots+l_j=i-j}j\langle\Lambda^{l_1}q,q\rangle_x\langle\Lambda^{l_2}q,q\rangle\cdots\langle\Lambda^{l_j}q,q\rangle.\end{aligned}$$

By using the following identities:

$$\sum_{i=1}^{k}\langle\Lambda^{k-i}q,q\rangle\sum_{j=1}^{i}a_jr_{i,j}=\sum_{j=1}^{k}a_j\sum_{i=0}^{k-j}\langle\Lambda^iq,q\rangle r_{k-i,j}, \tag{2.15a}$$

$$\begin{aligned}&\sum_{i=0}^{k}\langle\Lambda^iq,q\rangle\sum_{l_1+\cdots+l_j=k-i}\langle\Lambda^{l_1}q,q\rangle_x\langle\Lambda^{l_2}q,q\rangle\cdots\langle\Lambda^{l_j}q,q\rangle\\&=\frac{1}{j+1}D\sum_{l_1+\cdots+l_{j+1}=k}\langle\Lambda^{l_1}q,q\rangle\cdots\langle\Lambda^{l_{j+1}}q,q\rangle,\end{aligned} \tag{2.15b}$$

we obtain

$$\begin{aligned}u_{m-k}=&\langle\Lambda^{k-1}q,q\rangle+\sum_{i=1}^{k-1}\frac{-j-2}{2(j+1)}a_j\\&\times\sum_{l_1+\cdots+l_{j+1}=k-1-j}\langle\Lambda^{l_1}q,q\rangle\cdots\langle\Lambda^{l_{j+1}}q,q\rangle\end{aligned}$$

which leads to (2.14b) immediately.

III. THE HAMILTONIAN SYSTEM OBTAINED FROM THE EIGENVALUE PROBLEM

Substituting (2.14b) into (2.11) and using (2.15) gives

$$\begin{aligned}&p_{jx}=-\sum_{i=0}^{m}a_i\sum_{l_1+\cdots+l_{j+1}=m-i}\langle\Lambda^{l_1}q,q\rangle\cdots\langle\Lambda^{l_j}q,q\rangle\lambda_j^{l_{j+1}}q_j,\\&q_{jx}=p_j,\end{aligned} \tag{3.1a}$$

which can be written in canonical Hamiltonian system:

$$p_x=-\frac{\partial H_0}{\partial q},\quad q_x=\frac{\partial H_0}{\partial p}, \tag{3.1b}$$

where

$$p=(p_1,\ldots,p_N)^T\equiv(q_{1x},\ldots,q_{Nx})^T,$$

$$H_0=\frac{1}{2}\langle p,p\rangle+\frac{1}{2}\sum_{i=0}^{m}\tilde{b}_i$$

$$\times \sum_{l_1+\cdots+l_{i+1}=m-i} \langle \Lambda^{l_1} q,q\rangle \cdots \langle \Lambda^{l_{i+1}} q,q\rangle$$

$$\tilde{b}_0 = -1,$$

$$\tilde{b}_i = \frac{a_i}{(i+1)} = -\frac{1}{2}\tilde{b}_{i-1} = \frac{(-1)^{i-1}}{2^i}, \quad i = 0,1,2,\ldots . \tag{3.1c}$$

Assuming that (p,q) solves (3.1), we have from (2.12a)

$$DLu|_A = \sum_{j=1}^{N} \lambda_j \Psi_{jx},$$

which yields

$$Lu|_A = \sum_{j=1}^{N} \lambda_j \Psi_j + 2C_2 e_{m-1} + \sum_{i=0}^{m-2} \beta_i^{(1)} e_i,$$

where subscript A means to substitute (2.14) into the expression, and C_i, $\beta_i^{(k)}$ are constants. Using (2.12a) and (2.13b) repeatedly, we get

$$DL^k u|_A = \sum_{i=0}^{k} C_i \sum_{j=1}^{N} \lambda_j^{k-i} \Psi_{jx} \quad (C_0 = 1;\ C_1 = 0), \tag{3.2a}$$

$$L^k u|_A = \sum_{i=0}^{k} C_i \sum_{j=1}^{N} \lambda_j^{k-i} \Psi_j + 2C_{k+1} e_{m-1} + \sum_{i=0}^{m-2} \beta_i^{(k)} e_i. \tag{3.2b}$$

It is found from (3.2b) and (2.3b) immediately that

$$b_k|_A = \frac{1}{2}\langle \Lambda^{k-1} q,q\rangle + \frac{1}{2}\sum_{i=2}^{k-1} C_i \langle \Lambda^{k-1-i} q,q\rangle + C_k, \quad k \geqslant 1. \tag{3.3}$$

Observe that C_i are the constants of the motion for system (3.1). We now use (3.3) and the recursion formula (2.10) to calculate the independent constants of the motion for (3.1). First, by substituting (3.3) into (2.10a) we can show by induction that

$$C_2 = C_3 = \cdots = C_m = 0.$$

Inserting (3.3) into both sides of (2.10b) yields the following lemma.

Lemma:

$$C_{k+m} = F_{k+m} + \sum_{i+j+l=k+m} C_i C_j F_l + 2\sum_{i+j=k+m} C_i F_j - \frac{1}{2}\sum_{i+j=k+m} C_i C_j, \quad k = 1,2,\ldots, \tag{3.4}$$

where $i,j,l \geqslant 1$, $F_1 = F_2 = \cdots = F_m = 0$,

$$F_{k+m} = \frac{1}{4}\left[\langle \Lambda^{k-1} p,p\rangle + \sum_{i=0}^{m} \tilde{b}_i \sum_{l_1+\cdots+l_{i+1}=m-i} \langle \Lambda^{l_1} q,q\rangle \cdots \langle \Lambda^{l_i} q,q\rangle \langle \Lambda^{l_{i+1}+k-1} q,q\rangle + \frac{1}{2}\sum_{j=0}^{k-2} (\langle \Lambda^j p,p\rangle \langle \Lambda^{k-2-j} q,q\rangle - \langle \Lambda^j p,q\rangle \langle \Lambda^{k-2-j} p,q\rangle)\right], \quad k = 1,2,\ldots . \tag{3.5}$$

Proof: It is clear that in order to get F_{k+m} we just need to replace b_j $(j \geqslant 1)$ in (2.10b) by $\frac{1}{2}\langle \Lambda^{j-1} q,q\rangle$:

$$F_{k+m} = \frac{1}{8}\sum_{j=0}^{k-2}\Big[\underbrace{\langle \Lambda^j p_x,q\rangle}_{\langle 1\rangle} + \underbrace{\langle \Lambda^j p,p\rangle}_{\langle 2\rangle}\Big]\langle \Lambda^{k-2-j} q,q\rangle + \frac{1}{4}\Big[\underbrace{\langle \Lambda^{k-1} p_x,q\rangle}_{\langle 3\rangle} + \underbrace{\langle \Lambda^{k-1} p,p\rangle}_{\langle 4\rangle}\Big] - \frac{1}{8}\sum_{j=0}^{k-2}\langle \Lambda^j p,q\rangle \underbrace{\langle \Lambda^{k-2-j} p,q\rangle}_{\langle 5\rangle} + \frac{1}{8}\sum_{i=0}^{m-1} u_i \sum_{j=0}^{k+i-2} \langle \Lambda^j q,q\rangle \underbrace{\langle \Lambda^{k+i-2-j} q,q\rangle}_{\langle 6\rangle} + \frac{1}{2}\sum_{i=0}^{m-1} u_i \underbrace{\langle \Lambda^{k+i-1} q,q\rangle}_{\langle 7\rangle} - \frac{1}{8}\sum_{j=0}^{k+m-2} \langle \Lambda^j q,q\rangle \underbrace{\langle \Lambda^{k+m-2-j} q,q\rangle}_{\langle 8\rangle} - \frac{1}{2}\underbrace{\langle \Lambda^{k-1+m} q,q\rangle}_{\langle 9\rangle}. \tag{3.6a}$$

Then,

$$\langle 1\rangle \overset{(3.1)}{=\!=} -\frac{1}{8}\sum_{j=0}^{k-2}\sum_{i=0}^{m} a_i \sum_{l_1+\cdots+l_{i+1}=m-i} \langle \Lambda^{l_1} q,q\rangle \cdots \langle \Lambda^{l_i} q,q\rangle \langle \Lambda^{l_{i+1}+j} q,q\rangle \langle \Lambda^{k-2-j} q,q\rangle$$

$$\overset{(2.15a)}{=\!=} -\frac{1}{8}\sum_{i=0}^{m} a_i \sum_{j=0}^{k-2} \sum_{l_1+\cdots+l_{i+1}=m-i} \langle \Lambda^{l_1} q,q\rangle \cdots \langle \Lambda^{l_{i+1}+j} q,q\rangle \langle \Lambda^{k-2-j} q,q\rangle$$

$$\langle 6\rangle \overset{(2.14b)}{=\!=} \frac{1}{8}\sum_{i=0}^{m-1}\sum_{j=1}^{m-i} a_j \sum_{l_1+\cdots+l_j=m-i-j} \langle \Lambda^{l_1} q,q\rangle \cdots \langle \Lambda^{l_j} q,q\rangle \sum_{l=0}^{k+i-2} \langle \Lambda^l q,q\rangle \langle \Lambda^{k+i-2-l} q,q\rangle$$

$$\overset{(2.15a)}{=\!=} \frac{1}{8}\sum_{i=1}^{m} a_i \sum_{l_1+\cdots+l_{i+1}=m-i} \langle \Lambda^{l_1} q,q\rangle \cdots \langle \Lambda^{l_i} q,q\rangle \times\left[\sum_{l=0}^{k-3} \langle \Lambda^l q,q\rangle \langle \Lambda^{k-2+l_{i+1}-l} q,q\rangle + \sum_{l=k-2}^{k-2+l_{i+1}} \langle \Lambda^l q,q\rangle \langle \Lambda^{k-2+l_{i+1}-l} q,q\rangle\right]$$

$$= \frac{1}{8}\sum_{i=1}^{m} a_i \sum_{l_1+\cdots+l_{i+1}=m-i} \langle \Lambda^{l_1} q,q\rangle \cdots \langle \Lambda^{l_i} q,q\rangle \sum_{j=1}^{k-2} \langle \Lambda^{k-2-j} q,q\rangle \langle \Lambda^{l_{i+1}+j} q,q\rangle$$

$$+ \frac{1}{8}\sum_{i=1}^{m} a_i \sum_{l_1+\cdots+l_{i+2}=m-i} \langle \Lambda^{l_1} q,q\rangle \cdots \langle \Lambda^{l_{i+1}} q,q\rangle \langle \Lambda^{k-2+l_{i+2}} q,q\rangle,$$

$$\langle 1\rangle + \langle 6\rangle + \langle 8\rangle = \frac{1}{8}\sum_{i=1}^{m} a_{i-1} \sum_{l_1+\cdots+l_{i+1}=m-i} \langle \Lambda^{l_1} q,q\rangle \cdots \langle \Lambda^{l_i} q,q\rangle \langle \Lambda^{k-1+l_{i+1}} q,q\rangle. \tag{3.6b}$$

Similarly, we have

$$\langle 3\rangle + \langle 7\rangle + \langle 9\rangle = \frac{1}{4}\sum_{i=0}^{m} a_i \sum_{l_1+\cdots+l_{i+1}=m-i} \langle \Lambda^{l_1} q,q\rangle \cdots \langle \Lambda^{l_i} q,q\rangle \langle \Lambda^{k-1+l_{i+1}} q,q\rangle,$$

which together with (3.6a) and (3.6b) leads to (3.5).

In a similar way, we can show (3.4) by a straightforward calculation.

It is clear from (3.4) by induction that F_k defined by (3.5) are also the integrals of motion for system (3.1), that is, if (p,q) satisfy (3.1), then

$$\frac{dF_k}{dx} = 0. \tag{3.7a}$$

Also, a direct calculation shows that if (p,q) satisfies (3.1), then

$$\frac{\partial F_k}{\partial q} = -\frac{d}{dx}\frac{\partial F_k}{\partial p}. \tag{3.7b}$$

Set

$$G_k = \frac{1}{8}\sum_{j=0}^{k-2} (\langle \Lambda^j p,p\rangle \langle \Lambda^{k-2-j} q,q\rangle - \langle \Lambda^j p,q\rangle \langle \Lambda^{k-2-j} p,q\rangle),$$

$$Q_k = F_{k+m} - G_k.$$

It is known[1] that G_k are in involution with respect to the ordinary Poisson bracket defined as

$$\{f,g\} = \sum_{j=1}^{N}\left(\frac{\partial f}{\partial p_j}\frac{\partial g}{\partial q_j} - \frac{\partial f}{\partial q_j}\frac{\partial g}{\partial p_j}\right).$$

Using the identity

$$\sum_{i=0}^{l} \langle \Lambda^{l+k+j-i} p,p\rangle \langle \Lambda^i q,q\rangle + \sum_{i=0}^{k} \langle \Lambda^i p,p\rangle \langle \Lambda^{l+k+j-i} q,q\rangle$$

$$= \sum_{i=0}^{l+k+j} \langle \Lambda^{l+k+j-i} p,p\rangle \langle \Lambda^i q,q\rangle - \sum_{i=l+1}^{l+j-1} \langle \Lambda^{l+k+j-i} p,p\rangle \langle \Lambda^i q,q\rangle, \quad j = 1,2,...,$$

it can be shown by a straightforward calculation that

$$\{Q_k,G_l\} + \{G_k,Q_l\} + \{Q_k,Q_l\} = 0.$$

Thus integrals of motion F_k are in involution. Since all λ_k are distinct, the Vandermonde determinant of $\lambda_1,...,\lambda_N$ is not zero. Then it is easy to see that

$$\text{grad } F_k = \left(\frac{\partial F_k}{\partial q_1},...,\frac{\partial F_k}{\partial q_N},\frac{\partial F_k}{\partial p_1},...,\frac{\partial F_k}{\partial p_N}\right),$$

$k = m+1,...,m+N,$

are functionally independent. So we have the following propositions.

Proposition 2: The Hamiltonial system (3.1) is completely integrable in the sense of Liouville.[8]

Proposition 3: If (p,q) is a solution of (3.1), the u given by (2.14) satisfies a certain higher-order stationary equation:

$$DL^N u + \sum_{k=0}^{N-1} d_k\, DL^k u = 0, \tag{3.8}$$

where constants d_k are determined by $\lambda_1,...,\lambda_N$ and $C_1,...,C_N$.

Proof: Set

$$(\lambda-\lambda_1)\cdots(\lambda-\lambda_N) = \lambda^N + \sum_{k=1}^{N} g_k \lambda^{N-k}.$$

Using (3.2a) and (2.15a) yields

$$\sum_{k=0}^{N} d_k\, DL^k u|_A = \sum_{j=1}^{N} \Psi_{jx} \sum_{k=0}^{N} d_k \sum_{i=0}^{k} C_i \lambda_j^{k-i} = \sum_{j=1}^{N} \Psi_{jx} \sum_{k=0}^{N} \lambda_j^{N-k} \sum_{i=0}^{k} C_i d_{N-k+i}.$$

Take $d_N = 1$ and

$$d_{N-k} = g_k - \sum_{i=1}^{k} C_i d_{N-k+i}, \quad k = 1,...,N.$$

Then it is easy to see that (3.8) holds.

IV. THE HAMILTONIAN SYSTEMS OBTAINED FROM THE TIME PART

It is found from (3.4) by induction that

$$C_k = \sum_{j=1}^{k} c_j \sum_{m_1+\cdots+m_j=k} F_{m_1}\cdots F_{m_j}, \quad k = 1,2,..., \tag{4.1a}$$

where $m_1 \geqslant 1,...,m_j \geqslant 1$, $c_1 = 1$, $c_2 = 3/2$,

$$c_j = \sum_{l=1}^{j-2} c_l c_{j-l-1} + 2c_{j-1} - \frac{1}{2}\sum_{l=1}^{j-1} c_l c_{j-l}, \quad j = 2,3,... . \tag{4.1b}$$

We now consider systems obtained from (2.2):

$$\phi_{jt_n} = -\tfrac{1}{2}B_{jx}^{(n)}\phi_j + B_j^{(n)}\phi_{jx}, \quad B_j^{(n)} = B^{(n)}|_{\lambda=\lambda_j}, \quad j = 1,...,N. \tag{4.2}$$

Under the constraint condition (2.14) and (3.1), (4.2) becomes, by using (3.3), (3.5), and (4.1),

$$\phi_{jt_n} = \frac{1}{2}\sum_{k=1}^{n}\lambda_j^{n-k}\sum_{l=0}^{k-1}C_l(\langle\Lambda^{k-l-1}q,q\rangle p_j - \langle\Lambda^{k-l-1}p,q\rangle q_j) + \sum_{k=0}^{n}\lambda_j^{n-k}C_k p_j$$

$$= \frac{1}{2}\sum_{k=0}^{n-1}C_k\Bigg[\sum_{l=0}^{n-k-1}\langle\Lambda^{n-k-1-l}q,q\rangle\lambda_j^l p_j - \langle\Lambda^{n-k-l-1}p,q\rangle\lambda_j^l q_j + 2\lambda_j^{n-k}p_j\Bigg] + C_n p_j$$

$$= 2\sum_{k=0}^{n}C_k\frac{\partial F_{n-k+m+1}}{\partial p_j}$$

$$= 2\sum_{k+l=m+n+1}\left(\sum_{i=0}^{k}c_i\sum_{m_1+\cdots+m_i=k}F_{m_1}\cdots F_{m_i}\right)\times\frac{\partial F_l}{\partial p_j} \qquad (c_0=1, F_1=\cdots=F_m=0)$$

$$= 2\sum_{i=0}^{n}c_i\sum_{m_1+\cdots+m_{i+1}=n+m+1}F_{m_1}\cdots F_{m_i}\frac{\partial F_{m_{i+1}}}{\partial p_j}$$

$$= 2\frac{\partial}{\partial p_j}\left(\sum_{i=0}^{n}\frac{c_i}{i+1}\times\sum_{m_1+\cdots+m_{i+1}=n+m+1}F_{m_1}\cdots F_{m_{i+1}}\right) \qquad (m_j\geqslant 1),$$

which together with (3.7b) implies that (4.2) can be written in canonical Hamiltonian form:

$$q_{t_n} = \frac{\partial H_n}{\partial p}, \quad p_{t_n} = -\frac{\partial H_n}{\partial q}, \tag{4.3}$$

with

$$H_n = 2\sum_{i=0}^{n}\frac{c_i}{i+1}\sum_{m_1+\cdots+m_{i+1}=n+m+1}F_{m_1}\cdots F_{m_{i+1}}, \tag{4.4}$$

where $m_1\geqslant 1,\ldots,m_{i+1}\geqslant 1$, $F_1=\cdots=F_m=0$, $c_0=1$.

Proposition 4: The Hamiltonian systems (4.3) ($n=0,1,\ldots$, called $t_0=x$) are completely integrable and commute with each other. If (p,q) satisfies (3.1) and (4.3) (for fixed n, $n\geqslant 1$), then u given by (2.14) is a solution of Eq. (2.4).

Proof: Since F_k are in involution, it is clear from (4.4) that

$$\frac{dF_k}{dt_n} = \{H_n,F_k\} = 0, \quad \{H_n,H_k\} = 0, \quad n,k = 0,1,\ldots.$$

Thus systems (4.3) ($n=0,1,\ldots$) are completely integrable and commute with each other. Observe that (2.4) is deduced from the solvability condition of (2.11) and (4.2), (3.1) and (4.3) are obtained by substituting (2.14) into (2.11) and (4.2), respectively, it is easy to see that if (p,q) satisfies both (3.1) and (4.3) (for fixed n, $n\geqslant 1$), then u given by (2.14) solves Eq. (2.4).

ACKNOWLEDGMENTS

This work was supported by Scientific Research Fund of Academy of Sciences of China.

[1] J. Moser, "Integrable Hamiltonian systems and spectral theory," in *Proceedings of the 1983 Beijing Symposium on Differential Geometry and Differential Equations 1986,* edited by Liao Shantao (Science, Beijing, 1986), pp. 157–230.
[2] D. V. Choodnovsky, G. V. Choodnovsky, Nuovo Cimento B **40**, 399 (1977).
[3] H. Airault, H. P. Mckean, Jr., J. Moser, Commun. Pure Appl. Math **30**, 95 (1977).
[4] H. Flaschka, "Relations between infinite-dimensional and finite-dimensional isospectral equation," in *Proceedings of RIMS Symposium on Nonlinear Integrable Systems—Classical Theory and Quantum Theory,* Tyoto, 1981, edited by M. Jimbo and T. Miwa (World Scientific, Singapore, 1983), pp. 219–240.
[5] C. W. Cao, Chinese Q. J. Math. **3**(1), 90 (1988).
[6] Y. B. Zeng and Y. S. Li, J. Math. Phys. **30**, 1679 (1989).
[7] Y. B. Zeng and Y. S. Li, J. Phys. A: Math. Gen. **23**, L89 (1990).
[8] V. I. Arnold, *Mathematical Methods of Classical Mechanics* (MIR, Moscow, 1975) (English translation published by Springer, New York, 1978).
[9] L. M. Alonso, J. Math. Phys. **21**, 2342 (1980).
[10] A. S. Fokas and R. L. Anderson, J. Math. Phys. **23**, 1066 (1982).
[11] M. Antonowicz and A. P. Fordy, *Nonlinear Evolution Equations and Dynamical Systems* (NEEDS'87), edited by J. Leon (World Scientific, Singapore, 1988), pp. 145–160.

Introduction to a covariant theory of special functions of mathematical physics

J. C. Lucquiaud
Laboratoire de Physique Mathematique, College de France, 11, Place Marcelin Berthelot, 75231 Paris Cedex 05, France

(Received 29 March 1990; accepted for publication 27 June 1990)

Using harmonic analysis techniquies, the covariant expression of Jacobi and Hermite polynomials on an n-dimensional space endowed with a metric g of signature (p_+, q_-) is given. The properties of these polynomials are studied and their relations with the hypergeometric function are given.

I. INTRODUCTION

For the needs of relativistic physics, especially statistical mechanics, and the quantum theory of a field on a curved space-time, we found it useful to build, in a covariant way, the special functions of mathematical physics. This work is essentially devoted to Jacobi and Hermite polynomials on an n-dimensional space endowed with a metric g of signature (p_+, q_-).

We know indeed that a large number of special functions arise in a natural way in the groups representation theory through the zonal spherical functions on homogeneous spaces $\mathcal{M} = G/H$, G being an isometry group of $\mathcal{M}$ and H an isotropy compact subgroup of a point $\zeta \in \mathcal{M}$.

We will mainly use two properties of these zonal spherical functions: (a) They can be expressed with the Gauss hypergeometric function: They are Jacobi polynomials of order k in the compact case, $F(-k, k+2c-1; c; \sin^2\theta)$ and of order $k/2$ in the noncompact case,

$$F\left(-\frac{k}{2}, \frac{k+2c-1}{2}; c; -\mathrm{sh}^2\,\omega\right),$$

where θ and ω are, respectively, the Euler angles of Cartan and Iwasawa decompositions.[1]

(b) Up to a factor, the zonal spherical functions are, especially for the $SO(n)$ group, the values taken by the harmonic projection of $(x_n)^k$ on the isotropy sphere, x_n being the component of an $\mathbf{x}$ vector on the invariant vector by the isotropy group H.

We first notice that Hermite and Jacobi polynomials (therefore especially Gegembauer, Legendre, and Tchebycheff ones) of degree k may be written[2]

$$\mathscr{F}_k(x) = \sum_{p=0}^{[k/2]} F_{k-2p} x^{k-2P},$$

which lead us to look for a second-order differential equation admitting such polynomials as a particular solution. A simple calculation gives

$$(1-ux^2)\frac{d^2y}{dx^2} - vx\frac{dy}{dx} + k[v+(k-1)u]y = 0,$$

where u and v are two real parameters and $k \in N$.

The associated particular solution (polynomial) is

$$\mathscr{F}_k^{u,v}(x) = \sum_{p=0}^{[k/2]} \frac{(-1)^p}{2^{2p}p!u^p} \frac{\Gamma(k+1)}{\Gamma(k+1-2p)} \times \frac{\Gamma((v-u)/2u+k-p)}{\Gamma((v-u)/2u+k)} x^{k-2p},$$

which, up to a factor, gives us the classical polynomials for certain values of parameters u and v.

Examples:

Hermite polynomials: $H_k(x) = \mathscr{F}_k^{0,1}(x)$,
Tchebycheff polynomials: $T_k(x) = 2^{k-1}\mathscr{F}_k^{1,1}(x)$, etc.

Therefore, we reduce the analysis of classical polynomials to $\mathscr{F}_k^{u,v}(x)$ ones.[2] In order to get the covariant formulation of the classical polynomials, we have adapted Vilenkin's method[3] to the tensorial case.

With the help of the Laplace–Beltrami operator and the isomorphism between homogeneous k degree polynomials and same degree symmetric tensors, we first get canonical decomposition of the symmetric tensor. In a covariant way, zonal spherical functions are expressed with the harmonic projection $\tilde{x}^{\alpha_1\alpha_2\cdots\alpha_k}$ of $x^{\alpha_1\alpha_2\cdots\alpha_k} = x^{\alpha_1}x^{\alpha_2}\cdots x^{\alpha_k}$. The value of $\tilde{x}^{\alpha_1\alpha_2\cdots\alpha_k}$ on one of the basic vectors is nothing else than $\mathscr{F}_k^{1,n-1}(x)$, up to the sign (due to the metrics), then we get immediately the covariant formulation of classical polynomials on an n-dimensional space endowed with a metric of any given signature.[2]

Our work runs as follows.

(1) In Sec. II, we will analyze the previous differential equation and give the general solution in a canonical way so that it can be reduced to the classical polynomials of first and second kinds. The relations with the hypergeometric functions are also established.

(2) In Sec. III, we give the canonical decomposition of symmetric tensors from which we derive the covariant formulations of Sec. I polynomials.

(3) In Sec. IV, we give without proof (as they come immediately from the previous sections' results), the scalar and tensorial properties of classical polynomials together with the equations they satisfy.

 0022-2488/90/122840-11$03.00

II. ZONAL SPHERICAL FUNCTIONS

A. Fundamental equation

1. Preliminaries (Ref. 2)

Let us consider the following second-order differential equation

$$(1-ux^2)\frac{d^2y}{dx^2}-vx\frac{dy}{dx}+k\,[v+(k-1)u]y=0, \quad (1)$$

in which $k\in N$ and u and v are two real parameters such that

$$v+(2k-2p-1)\,u\neq 0, \quad \text{when } 1\leqslant p\leqslant [k/2]. \quad (2)$$

In particular, this leads to u and v not equal to zero at the same time.

With such a condition, Eq. (1) has a particular polynomial solution with the following form:

$$y=\sum_{p=0}^{[k/2]} f_{k-2p}x^{k-2p}.$$

Coefficients f_{k-2p} satisfy the recursion formula

$$(k-2p+1)(k-2p+2)\,f_{k-2p+2} = -2p[v+(2k-2p-1)u]\,f_{k-2p},$$

$$\text{where } 1\leqslant p\leqslant [k/2]. \quad (3)$$

We define polynomials $\mathcal{F}_k^{u,v}(x)$, particularly the solution of (1) by setting $f_k=1$. Therefore,

$$\mathcal{F}_k^{u,v}(x)=\sum_{p=0}^{[k/2]}\frac{(-1)^p}{2^{2p}p!u^p}\frac{\Gamma(k+1)}{\Gamma(k+1-2p)} \times\frac{\Gamma[(v-u)/2u+k-p]}{\Gamma[(v-u)/2u+k]}x^{k-2p}, \quad (4)$$

Note: If $v\neq 0$, we have by continuity

$$\lim_{u\to 0}\mathcal{F}_k^{u,v}(x)=\mathcal{F}_k^{0,v}(x) =\sum_{p=0}^{[k/2]}\frac{(-1)^p}{2^p p!v^p}\frac{\Gamma(k+1)}{\Gamma(k+1-2p)}x^{k-2p}. \quad (5)$$

Examples: (a) If $u=0, v=1$, (1) becomes the Hermite equation:

$$\frac{d^2y}{dx^2}-x\frac{dy}{dx}+ky=0, \quad (6)$$

and Hermite polynomials are exactly

$$H_k(x)=\mathcal{F}_k^{0,1}(x) =\sum_{p=0}^{[k/2]}\frac{(-1)^p}{2^p p!}\frac{\Gamma(k+1)}{\Gamma(k+1-2p)}x^{k-2p}. \quad (7)$$

(b) If $u=1, v=2$, we recognize the Legendre equation

$$(1-x^2)\frac{d^2y}{dx^2}-2x\frac{dy}{dx}+k(k+1)y=0, \quad (8)$$

with, consequently, the following particular solution:

$$\mathcal{F}_k^{1,2}(x)=\sum_{p=0}^{[k/2]}\frac{(-1)^p}{2^{2p}p!}\frac{\Gamma(k+1)}{\Gamma(k+1-2p)} \times\frac{\Gamma(\frac{1}{2}+k-p)}{\Gamma(\frac{1}{2}+k)}x^{k-2p}. \quad (9)$$

2. General solution

Let us now look for the general solution of Eq. (1) in a series expansion:

$$y=\sum_{p=0}^{\infty}a_p x^p.$$

We immediately get the recursion formula

$$(p+1)(p+2)a_{p+2} = (p-k)[v+(k-1+p)u]a_p \text{ with } p\geqslant 0, \quad (10)$$

Results are: (a) If $u\neq 0$, the general solution of the equation

$$(1-ux^2)\frac{d^2y}{dx^2}-vx\frac{dy}{dx}+k\,[v+(k-1)u]y=0, \quad (11)$$

is given by

$$y=a_0F\left(-\frac{k}{2},\frac{v+(k-1)u}{2u};\frac{1}{2};ux^2\right) +a_1xF\left(\frac{1-k}{2},\frac{v+ku}{2u};\frac{3}{2};ux^2\right),$$

in which $F(a,b,c;t)$ is the Gauss hypergeometric function:

$$F(a,b;c;t)=\sum_{p=0}^{\infty}\frac{\Gamma(a+p)}{\Gamma(a)}\frac{\Gamma(b+p)}{\Gamma(b)}\frac{\Gamma(c)}{\Gamma(c+p)}\frac{t^p}{p!} \quad (12)$$

which is the solution of the differential equation

$$t(1-t)\frac{d^2y}{dt^2}+[c-(a+b+1)t]\frac{dy}{dt}-aby=0. \quad (13)$$

(b) If $v\neq 0$, the general solution of the equation

$$\frac{d^2y}{dx^2}-vx\frac{dy}{dx}+kvy=0$$

is given by (14)

$$y=a_0\Phi\left(-\frac{k}{2};\frac{1}{2};\frac{vx^2}{2}\right)+a_1x\Phi\left(\frac{1-k}{2};\frac{3}{2};\frac{vx^2}{2}\right),$$

in which $\Phi(a;c;t)$ is the Kummer function (confluent hypergeometric function):

$$\Phi(a;c;t)=\sum_{p=0}^{\infty}\frac{\Gamma(a+p)}{\Gamma(a)}\frac{\Gamma(c)}{\Gamma(c+p)}\frac{t^p}{p!}, \quad (15)$$

solution of the equation

$$\left[t\frac{d^2y}{dt^2}+(c-t)\right]\frac{dy}{dt}-ay=0. \quad (16)$$

3. First- and second-class solutions

(a) We notice that if k is even,

$$F\left(-\frac{k}{2},\frac{v+(k-1)u}{2u};\frac{1}{2};ux^2\right)$$

is a polynomial with the following form

$$F\left(-\frac{k}{2},\frac{v+(k-1)u}{2u};\frac{1}{2};ux^2\right)=\sum_{p=0}^{[k/2]}a_{k-2p}\,x^{k-2p}.$$

Using (10) we immediately verify that coefficients a_{k-2p} satisfy the same recursion formula (3) as coefficients of polynomial $\mathcal{F}_k^{u,v}(x)$. This leads to

$$F\left(-k, \frac{v-u}{2u}+k; \frac{1}{2}; ux^2\right)$$
$$= (-u)^k \frac{\Gamma[(v-u)/2u+2k]\Gamma(\frac{1}{2})}{\Gamma[(v-u)/2u+k]\Gamma(\frac{1}{2}+k)} \mathcal{F}_{2k}^{u,v}(x)$$

and in the same way (17)

$$xF\left(-k, \frac{v+u}{2u}+k; \frac{3}{2}; ux^2\right)$$
$$= (-u)^k \frac{\Gamma[(v+u)/2u+2k]\Gamma(\frac{3}{2})}{\Gamma[(v+u)/2u+k]\Gamma(\frac{3}{2}+k)} \mathcal{F}_{2k+1}^{u,v}(x).$$

From that, the general solution of the equation

$$(1-ux^2)\frac{d^2y}{dx^2} - vx\frac{dy}{dx} + k[v+(k-1)u]y = 0$$

may be written, (18)

$$y = a\mathcal{F}_k^{u,v}(x) + b\mathcal{G}_k^{u,v}(x)$$

in which $\mathcal{F}_k^{u,v}(x)$ is the polynomial defined in (4), and $\mathcal{G}_k^{u,v}(x)$ is a nonpolynomial defined by

$$\mathcal{G}_{2k}^{u,k}(x) = xF\left(\frac{1}{2}-k, \frac{v}{2u}+k; \frac{3}{2}; ux^2\right),$$
$$\mathcal{G}_{2k+1}^{u,v}(x) = F\left(-\frac{1}{2}-k, \frac{v}{2u}+k; \frac{1}{2}; ux^2\right). \quad (19)$$

(b) If $v \neq 0$, we get the relations

$$\Phi\left(-k; \frac{1}{2}; \frac{vx^2}{2}\right) = (-2v)^k \frac{\Gamma(1+k)}{\Gamma(1+2k)} \mathcal{F}_{2k}^{0,v}(x),$$
$$x\Phi\left(-k; \frac{3}{2}; \frac{vx^2}{2}\right) = (-2v)^k \frac{\Gamma(1+k)}{\Gamma(2+2k)} \mathcal{F}_{2k+1}^{0,v}(x). \quad (20)$$

Then the general solution of the equation

$$\frac{d^2y}{dx^2} - vx\frac{dy}{dx} + kvy = 0$$

may be written, (21)

$$y = a\mathcal{F}_k^{0,v}(x) + b\mathcal{G}_k^{0,v}(x),$$

in which $\mathcal{F}_k^{0,v}(x)$ is the polynomial defined in (5), and $\mathcal{G}_k^{0,v}(x)$ is a nonpolynomial defined by

$$\mathcal{G}_{2k}^{0,v}(x) = x\Phi\left(\frac{1}{2}-k; \frac{3}{2}; \frac{vx^2}{2}\right),$$
$$\mathcal{G}_{2k+1}^{0,v}(x) = \Phi\left(-\frac{1}{2}-k; \frac{1}{2}; \frac{vx^2}{2}\right). \quad (22)$$

The general solution (18) of the particular case (21) is canonical in the following way.

Let $y(x)$ be a particular solution of Eq. (1). If it is polynomial, it can be brought down to $\mathcal{F}_k^{u,v}(x)$, up to a given factor, and in the contrary case to $\mathcal{G}_k^{u,v}(x)$. We will name $\mathcal{F}_k^{u,v}(x)$ the first-class solution and $\mathcal{G}_k^{u,v}(x)$ the second-class solution.

4. Application to the Hermite equation

From previous results, the Hermite equation

$$\frac{d^2y}{dx^2} - x\frac{dy}{dx} + ky = 0$$

admits as a general solution (23)

$$y = a\mathcal{F}_k^{0,1}(x) + b\mathcal{G}_k^{0,1}(x).$$

Therefore, (a) Hermite polynomials are shown as

$$H_k(x) = \mathcal{F}_k^{0,1}(x), \quad (24)$$

and we have

$$H_{2k}(x) = \mathcal{F}_{2k}^{0,1}(x) = \left(-\frac{1}{2}\right)^k \frac{\Gamma(2k+1)}{\Gamma(k+1)} \times \Phi\left(-k; \frac{1}{2}; \frac{x^2}{2}\right),$$
$$H_{2k+1}(x) = \mathcal{F}_{2k+1}^{0,1}(x) = \left(-\frac{1}{2}\right)^k \frac{\Gamma(2k+2)}{\Gamma(k+1)} \times x\Phi\left(-k; \frac{3}{2}; \frac{x^2}{2}\right). \quad (25)$$

(b) Hermite functions are

$$h_{2k}(x) = (-2)^k\Gamma(k+1)\mathcal{G}_{2k}^{0,1}(x)$$
$$= (-2)^k\Gamma(k+1)x\Phi\left(\frac{1}{2}-k; \frac{3}{2}; \frac{x^2}{2}\right),$$
$$h_{2k+1}(x) = -(-2)^k\Gamma(k+1)\mathcal{G}_{2k+1}^{0,1}(x)$$
$$= -(-2)^k\Gamma(k+1)\Phi\left(-\frac{1}{2}-k; \frac{1}{2}; \frac{x^2}{2}\right). \quad (26)$$

B. Zonal spherical functions

Let us consider below Eq. (1) with $u = 1$ and $v = n-1$, n integer $\geqslant 2$. We will set

$$Z_k^{(n)}(x) = \mathcal{F}_k^{1,n-1}(x), \quad \mathcal{S}_k^{(n)}(x) = \mathcal{G}_k^{1,n-1}(x), \quad (27)$$

leading to the following result.

The equation

$$(1-x^2)\frac{d^2y}{dx^2} - (n-1)x\frac{dy}{dx} + k(k+n-2)y = 0$$

admits as a general solution (28)

$$y = aZ_k^{(n)}(x) + b\mathcal{S}_k^{(n)}(x).$$

From (4), the first solution is

$$Z_k^{(n)}(x) = \sum_{p=0}^{[k/2]} \frac{(-1)^p}{2^{2p}p!} \frac{\Gamma(k+1)}{\Gamma(k+1-2p)} \times \frac{\Gamma[(n-2)/2+k-p]}{\Gamma[(n-2)/2+k]} x^{k-2p}, \quad (29)$$

and with (17) we get

$$F\left(-k, \frac{n-2}{2}+k; \frac{1}{2}; x^2\right)$$
$$= (-1)^k \frac{\Gamma[(n-2)/2+2k]\Gamma(\frac{1}{2})}{\Gamma[(n-2)/2+k]\Gamma(\frac{1}{2}+k)} Z_{2k}^{(n)}(x),$$
$$xF\left(-k, \frac{n}{2}+k; \frac{3}{2}; x^2\right)$$
$$= (-1)^k \frac{\Gamma(n/2+2k)\Gamma(\frac{3}{2})}{\Gamma(n/2+k)\Gamma(\frac{3}{2}+k)} Z_{2k+1}^{(n)}(x). \quad (30)$$

The second-class solution is given by

$$\mathscr{S}^{(n)}_{2k}(x) = xF\left(\frac{1}{2}-k;\frac{n-1}{2}+k;\frac{3}{2};x^2\right),$$
$$\mathscr{S}^{(n)}_{2k+1}(x) = F\left(-\frac{1}{2}-k;\frac{n-1}{2}+k;\frac{1}{2};x^2\right). \quad (31)$$

C. Other relations with the hypergeometric function

(a) In the hypergeometric function (13), let us take $t=(1+\epsilon x)/2$, where $\epsilon = \pm 1$. It becomes

$$(1-x^2)\frac{d^2y}{dx^2} - [(a+b+1)x + (a+b+1-2c)\epsilon] \times\frac{dy}{dx} - aby = 0,$$

which will be Eq. (28) if $a=-k, b=k+n-2$; $c=(n-1)/2$. We then infer that

$$F\left(-k,k+n-2;\frac{n-1}{2};\frac{1+\epsilon x}{2}\right) = \left(-\frac{\epsilon}{2}\right)^k \frac{\Gamma(2k+n-2)\Gamma[(n-1)/2]}{\Gamma(k+n-2)\Gamma[(n-1)/2+k]} \times Z^{(n)}_k(x) \quad (\text{if } n\geqslant 2) \quad (32)$$
$$= (-2\epsilon)^k \frac{\Gamma[(n-2)/2+k]\Gamma(n-2)}{\Gamma[(n-2)/2]\Gamma(n-2+k)} \times Z^{(n)}_k(x) \quad (\text{if } n\geqslant 3).$$

(b) In a similar way, taking $t=(1-x^2)$ in (13), we get the general solution (28) in the following form:

$$y = aF\left(-\frac{k}{2},\frac{k+n-2}{2};\frac{n-1}{2};1-x^2\right) + bxF\left(\frac{1-k}{2},\frac{k+n-1}{2};\frac{n-1}{2};1-x^2\right), \quad (33)$$

which gives, in particular, for the first-class solution (polynomial),

$$F\left(-k,\frac{n-2}{2}+k;\frac{n-1}{2};1-x^2\right) = \frac{\Gamma[(n-2)/2+2k]\Gamma[(n-1)/2]}{\Gamma[(n-2)/2+k]\Gamma[(n-1)/2+k]} Z^{(n)}_{2k}(x),$$
$$xF\left(-k,\frac{n}{2}+k;\frac{n-1}{2};1-x^2\right) = \frac{\Gamma(n/2)+2k)\Gamma[(n-1)/2]}{\Gamma(n/2+k)\Gamma[(n-1)/2+k]} Z^{(n)}_{2k+1}(x). \quad (34)$$

D. Application to classical polynomials

Using harmonic analysis technics on homogeneous spaces, we get the zonal spherical functions in terms of the hypergeometric functions.[1]

In the compact case, the zonal functions are Jacobi polynomials of order k:

$$J_k(x) = F[-k,k+n-2;(n-1)/2;(1-x)/2] \quad (35)$$

and in the noncompact case, Jacobi polynomicals of order $k/2$:

$$J_{k/2}(x) = F\left(-\frac{k}{2},\frac{k+n-2}{2};\frac{n-1}{2};1-x^2\right). \quad (36)$$

From the previous results, it appears that all these functions may be brought down to first-class functions (29) or second-class ones (31).

E. Examples

1. Gegembauer polynomials

On an n-dimensional space ($n\geqslant 3$), they are defined by[4]

$$C^{(n-2)/2}_k(x) = \frac{(-\epsilon)^k\Gamma(k+n-2)}{\Gamma(k+1)\Gamma(n-2)} \times F\left(-k,n-2+k;\frac{n-1}{2};\frac{1+\epsilon x}{2}\right) \quad (\epsilon=\pm 1), \quad (37)$$

from which we get, by use of the preceding transformations,

$$C^{(n-2)/2}_k(x) = \frac{2^k\Gamma[(n-2)/2+k]}{\Gamma(k+1)\Gamma[(n-2)/2]} Z^{(n)}_n(x), \quad (38)$$
$$C^{(n-2)/2}_{2k}(x) = \frac{\Gamma(2k+n-2)}{\Gamma(2k+1)\Gamma(n-2)} \times F\left(-k,\frac{n-2}{2}+k;\frac{n-1}{2};1-x^2\right),$$
$$C^{(n-2)/2}_{2k+1}(x) = \frac{\Gamma(2k+n-1)}{\Gamma(2k+2)\Gamma(n-2)} \times xF\left(-k,\frac{n}{2}+k;\frac{n-1}{2};1-x^2\right), \quad (39)$$
$$C^{(n-2)/2}_{2k}(x) = \frac{(-1)^k\Gamma[(n-2)/2+k]}{\Gamma(k+1)\Gamma[(n-2)/2]} \times F\left(-k,\frac{n-2}{2}+k;\frac{1}{2};x^2\right),$$
$$C^{(n-2)/2}_{2k+1}(x) = \frac{(-1)^k(2k+n-2)\Gamma[(n-2)/2+k]}{\Gamma(k+1)\Gamma[(n-2)/2]} \times xF\left(-k,\frac{n}{2}+k;\frac{3}{2};x^2\right). \quad (40)$$

2. Legendre polynomials

Legendre polynomials, $P_x(x)$ are the Gegembauer polynomials when $n=3$. Then

$$P_k(x) = C^{1/2}_k(x). \quad (41)$$

3. Tchebycheff polynomials

Tchebycheff polynomials correspond to the case when $n=2$, and are defined by[4]

$$T_k(x) = (-\epsilon)^k F(-k,k;\tfrac{1}{2};[1+\epsilon x/2]) \quad (\epsilon=\pm 1), \quad (42)$$

then

$$T_k(x) = 2^{k-1}Z^{(2)}_k(x) \quad (k\geqslant 1), \quad (43)$$
$$T_{2k}(x) = F(-k,k;\tfrac{1}{2};1-x^2) = (-1)^k F(-k,k;\tfrac{1}{2};x^2),$$
$$T_{2k+1}(x) = xF(-k,k+1;\tfrac{1}{2};1-x^2) = (-1)^k(2k+1)xF(-k,k+1;\tfrac{3}{2};x^2). \quad (44)$$

F. Second-class functions case

For the corresponding second-class functions, we can get similar relations. They will be detailed in a later paper.

In the case of the Tchebycheff functions (of second class); $t_k(x)$, they are linked to the first-class polynomials $T_k(x)$ by

$$t_k(x)=\frac{1}{k}\sqrt{1-x^2}\,\frac{dT_k(x)}{dx}\quad(k\neq 0). \tag{45}$$

III. GEOMETRICAL INTERPRETATION

A. Canonical decomposition of symmetric tensors (Ref. 2)

1. Definitions

(a) We consider the n-dimensional space E_n (n finite) endowed with a metric of signature (p_+,q_-) with $p+q=n$. Afterwards, only Green indices will have the tensorial character. For the metric tensor, we have

$$g_{\alpha\beta}\text{ and }g^{\alpha\beta}=\begin{cases}0, & \text{if } \alpha\neq\beta,\\ +1, & \text{if } \alpha=\beta=1,2,\ldots,p,\\ -1, & \text{if } \alpha=\beta=p+1,p+2,\ldots,n\end{cases} \tag{46}$$

and

$$g_{\alpha\lambda}g^{\beta\lambda}=\delta_\alpha^\beta\quad\text{and}\quad g_{\alpha\beta}g^{\alpha\beta}=n, \tag{47}$$

where δ_α^β stands for the Kronecker symbol.

Here, E_n will be referred to as a reference frame $\{\mathbf{e}_\alpha\}$ such that $(\mathbf{e}_\alpha,\mathbf{e}_\beta)=g_{\alpha\beta}$. For all $\mathbf{x}\in E_n$, we then have the fundamental quadratic form

$$r^2=g_{\alpha\beta}x^\alpha x^\beta.$$

(b) Let $\mathscr{K}_k^{(n)}(\mathbf{x})$ be the space of homogeneous polynomials of degree k with n variables $(x^1,x^2,\ldots,x^n)$ components of a vector $\mathbf{x}\in E_n$.

We know that $\mathscr{K}_k^{(n)}(x)$ is isomorphic to the space $SE_n^{\otimes k}$ of symmetric tensors built on E_n. Therefore,

$$\text{dimension of }\mathscr{K}_k^{(n)}(\mathbf{x})=\text{dimension of }SE^{\otimes k}=\Gamma(k+n)/\Gamma(k+1)\Gamma(n).$$

If $f_k(\mathbf{x})\in\mathscr{K}_k^{(n)}(x)$, we will denote $f^{\alpha_1\alpha_2\cdots\alpha_k}$ as the corresponding symmetric tensor.

If $\mathbf{x}\in E_n$, we define

$$X^{\alpha_1\alpha_2\cdots\alpha_k}=x^{\alpha_1}x^{\alpha_2}\cdots x^{\alpha_k}.$$

If P and Q are two symmetric tensors, we note $[P_p,Q_q]^{\alpha_1\alpha_2\cdots\alpha_{p+q}}$ is their symmetric product; and if l is an integer, $[P_p,lg]^{\alpha_1\alpha_2\cdots\alpha_{p+2l}}$ represents the tensor obtained by the symmetric product between the tensor P of order p and l and the time metric tensor g. Particularly, $[X_k,lg]^{\alpha_1\alpha_2\cdots\alpha_{k+2l}}$ is, up to a change of notation, the Pichon–Marie tensor.

We will take

$$[X_0,0g]=1\quad\text{and}\quad[X_{k-2l},lg]=0,\quad\text{if } k<2l.$$

(c) Let $\widetilde{\mathscr{K}}_k^{(n)}(\mathbf{x})$ be the space of homogenous and harmonic polynomials of degree k with n variables; $\widetilde{\mathscr{K}}_k^{(n)}(\mathbf{x})$ is a subspace of $\mathscr{K}_k^{(n)}(\mathbf{x})$, and is isomorphic to the space $\widetilde{SE}_n^{\otimes k}$ of symmetric and harmonic tensors. We have

$$\text{dimension of }\widetilde{\mathscr{K}}_k^{(n)}(\mathbf{x})=\text{dimension of }\widetilde{SE}_n^{\otimes k}=\frac{(n+2k-2)\Gamma(n+k-2)}{\Gamma(k+1)\Gamma(n-1)}.$$

If $\tilde{q}^{\alpha_1\alpha_2\cdots\alpha_k}\in\widetilde{SE}_n^{\otimes k}$, we have $\Delta\tilde{q}^{\alpha_1\alpha_2\cdots\alpha_k}=0$, Δ being the Laplace–Beltrami operator, $\Delta=g^{\alpha\beta}\partial_{\alpha\beta}$, with

$$\partial_\alpha=\frac{\partial}{\partial x^\alpha},\quad\partial_{\alpha\beta}=\frac{\partial^2}{\partial x^\alpha\,\partial x^\beta},\text{etc.}$$

To simplify writing, we take

$$\partial^\alpha=g^{\alpha\lambda}\partial_\lambda,\quad\partial^{\alpha\beta}=g^{\alpha\lambda}g^{\beta\mu}\partial_{\lambda\mu},\text{etc.}$$

2. Useful relations

Let us indicate now some relations are useful afterwards.

a. Euler relations: For the symmetric tensor $X^{\alpha_1\alpha_2\cdots\alpha_k}=x^{\alpha_1}x^{\alpha_2}\cdots x^{\alpha_k}$, it is easy to get the Euler relation

$$x^\lambda\partial_\lambda(X^{\alpha_1\alpha_2\cdots\alpha_k})=kX^{\alpha_1\alpha_2\cdots\alpha_k}, \tag{48}$$

that we can generalize to

$$x^{\lambda_1}x^{\lambda_2}\cdots x^{\lambda_t}\partial_{\lambda_1\lambda_2\cdots\lambda_t}(X^{\alpha_1\alpha_2\cdots\alpha_k})=[\Gamma(k+1)/\Gamma(k+1-t)]X^{\alpha_1\alpha_2\cdots\alpha_k}.$$

Then we define the Euler operator D_t of order t by

$$D_t=x^{\lambda_1}x^{\lambda_2}\cdots x^{\lambda_t}\partial_{\lambda_1\lambda_2\cdots\lambda_t}, \tag{49}$$

so that

$$D_t(X^{\alpha_1\alpha_2\cdots\alpha_k})=[\Gamma(k+1)/\Gamma(k+1-t)]X^{\alpha_1\alpha_2\cdots\alpha_k}. \tag{50}$$

b. Expression of $\partial_{\lambda_1\lambda_2\cdots\lambda_s}(r^{2t})$*:* With $r^2=g_{\alpha\beta}x^\alpha x^\beta$, we have $\partial_\lambda(r^{2t})=2tr^{2t-2}x_\lambda$, and by recurrence we get

$$\partial_{\lambda_1\lambda_2\cdots\lambda_s}(r^{2t})=\Gamma(t+1)\sum_{p=0}^{[s/2]}\frac{2^{s-p}r^{2t-2s+2p}}{\Gamma(t+1-s+p)}\times[X_{s-2p},pg]_{\lambda_1\lambda_2\cdots\lambda_s}. \tag{51}$$

c. Expression of $\Delta^m(2^{2t})$*:* With the previous relation, we first have

$$\partial_{\lambda_1\lambda_2}(r^{2t})=2t(2t-2)r^{2t-4}X_{\lambda_1\lambda_2}+2tr^{2t-2}g_{\lambda_1\lambda_2}$$

and

$$\Delta(r^{2t})=2t(2t+n-2)r^{2t-2} \tag{52}$$

(we remind the reader that n is the dimension of the space).

By recurrence, we end up with

$$\Delta^m(r^{2t})=\frac{2^{2m}\Gamma(t+1)\Gamma(n/2)+t)}{\Gamma(t+1-m)\Gamma[(n/2)+t-m]}r^{2t-2m}. \tag{53}$$

d. Expression of $\Delta^m[r^{2t}f_k(\mathbf{x})]$ *with* $f_k(\mathbf{x})\in\mathscr{K}_k^{(n)}(\mathbf{x})$*:* If $f_k(\mathbf{x})\in\mathscr{K}_k^{(n)}(x)$, we first have

$$\Delta[r^{2t}f_k(\mathbf{x})]=\Delta(r^{2t})f_k(\mathbf{x})+2\partial^\lambda(r^{2t})\partial_\lambda[f_k(\mathbf{x})]+r^{2t}\Delta[f_k(\mathbf{x})].$$

Using (53) and the Euler relation (50), we get by recurrence:

$$\Delta^m[r^{2t}f_k(\mathbf{x})]=\sum_{p=0}^{m}\frac{2^{2p}\Gamma(m+1)\Gamma(t+1)\Gamma(a-m+p)}{\Gamma(p+1)\Gamma(m+1-p)\Gamma(t+1-p)\Gamma(a-m)}r^{2t-2p}\Delta^{m-p}[f_k(\mathbf{x})] \tag{54}$$

with $a = n/2 + k + t$.

e. Expression of $\Delta^m(RQ)$*:* If R and Q are two symmetric tensors (a zero-order tensor being considered as a scalar), we have

$$\Delta(RQ) = (\Delta R)Q + 2(\partial^\lambda R)(\partial_\lambda Q) + R(\Delta Q).$$

By recurrence and using the Leibnitz formula for a product we have

$$\Delta^m(RQ) = \Gamma(m+1)\sum_{p=0}^{m}\sum_{t=0}^{m-P}\frac{2^P}{\Gamma(P+1)\Gamma(t+1)\Gamma(m+1-P-t)}\partial_{\lambda_1\cdots\lambda_1}(\Delta^{m-p-t}R)\partial^{\lambda_1\cdots\lambda_p}(\Delta^t Q). \tag{55}$$

f. Expression of $\Delta^m(x^{\alpha_1\alpha_2\cdots\alpha_k})$*:* With the help of the Laplace–Beltrami operator $\Delta = g^{\alpha\beta}\partial_{\alpha\beta}$, we easily get for the symmetric tensor $x^{\alpha_1\alpha_2\cdots\alpha_k} = x^{\alpha_1}x^{\alpha_2\cdots\alpha_k}$:

$$\Delta^m(x^{\alpha_1\alpha_2\cdots\alpha_k}) = 2^m\Gamma(m+1)[X_{k-2m},mg]^{\alpha_1\alpha_2\cdots\alpha_k}, \tag{56}$$

which gives the meaning of the tensor $[X_{k-2m},mg]^{\alpha_1\alpha_2\cdots\alpha_k}$ which was introduced by Pichon–Marle during the calculation of moments of the distribution function of a relativistic fluid.[5]

B. Remarks

(a) Relation (56) is not dependent on the dimension n of the space. Let us then take $n = 1$. Assuming $g_{11} = +1$, we have

$$\frac{d^{2m}(x^k)}{dx^{2m}} = \frac{\Gamma(k+1)}{\Gamma(k+1-2m)}x^{k-2m} = 2^m\Gamma(m+1)Ax^{k-2m},$$

where A is the number of terms of the tensor $[x_{k-2m},mg]^{\alpha_1\alpha_2\cdots\alpha_k}$. We deduce

$$A = \frac{1}{2^m}\frac{\Gamma(k+1)}{\Gamma(m+1)\Gamma(k+1-2m)}. \tag{57}$$

(b) In the same way, for $n = 1$ and $r^2 = x^2$, relation (53) gives

$$\frac{\Gamma(1+2t)}{\Gamma(1+2t-2m)} = 2^{2m}\frac{\Gamma(t+1)\Gamma(\frac{1}{2}+t)}{\Gamma(1+t-m)\Gamma(\frac{1}{2}+t-m)}. \tag{58}$$

Also, relations (51), (54), and (56) lead immediately to other properties of the gamma function.

C. Canonical decomposition of symmetric tensors

We know that the space $\mathscr{K}_k^{(n)}(\mathbf{x})$ of homogeneous polynomials of degree k with n variables is a direct sum of the space $\tilde{\mathscr{K}}_k^{(n)}(\mathbf{x})$ of homogeneous and harmonic polynomials of degree k with n variables, and of the space $r^2\mathscr{K}_{k-2}^{(n)}(\mathbf{x})$ of $r^2 f_{k-2}(\mathbf{x})$ form polynomials when $f_{k-2}(\mathbf{x})\in\mathscr{K}_{k-2}^{(n)}(\mathbf{x})$:

$$\mathscr{K}_k^{(n)}(\mathbf{x}) = \tilde{\mathscr{K}}_k^{(n)}(\mathbf{x}) \oplus r^2\mathscr{K}_{k-2}^{(n)}(\mathbf{x}),$$

so that all polynomials $f_k(\mathbf{x})\in\mathscr{K}_k^{(n)}(\mathbf{x})$ have a unique decomposition in the following form:

$$f_k(\mathbf{x}) = \tilde{f}_k(\mathbf{x}) + r^2 f_{k-2}(\mathbf{x}),$$

where

$$\tilde{f}_k(\mathbf{x})\in\tilde{\mathscr{K}}_k^{(n)}(\mathbf{x})$$

and

$$f_{k-2}(\mathbf{x})\in\mathscr{K}_{k-2}^{(n)}(\mathbf{x})$$

and where $\tilde{f}_k(\mathbf{x})$ is the harmonic projection of $f_k(\mathbf{x})$.

Continuing the previous decomposition, we get

$$\mathscr{K}_k^{(n)}(\mathbf{x}) = \bigoplus_{p=0}^{[k/2]} r^{2p}\tilde{\mathscr{K}}_{k-2p}^{(n)}, \tag{59}$$

and consequently, all homogeneous polynomials $f_k(\mathbf{x})$ have a unique decomposition:

$$f_k(\mathbf{x}) = \sum_{p=0}^{[k/2]} r^{2p}\tilde{f}_{k-2p}(\mathbf{x}),$$

where

$$\tilde{f}_{k-2p}(\mathbf{x})\in\tilde{\mathscr{K}}_{k-2p}^{(n)}(\mathbf{x}). \tag{60}$$

1. Canonical decomposition of homogeneous polynomials

To obtain the harmonic components $\tilde{f}_{k-2p}(\mathbf{x})$ of $f_k(\mathbf{x})$, let us apply m times the Laplace operator on the two members of (60):

$$\Delta^m[f_k(\mathbf{x})] = \sum_{p=0}^{[k/2]}\Delta^m[r^{2p}\tilde{f}_{k-2p}(\mathbf{x})].$$

Using (54) and the fact that $\tilde{f}_{k-2p}(\mathbf{x})$ is harmonic, we get

$$\Delta^m[r^{2p}\tilde{f}_{k-2p}(\mathbf{x})] = \frac{2^{2m}\Gamma[p+1)\Gamma(n/2)+k-p]}{\Gamma(p+1-m)\Gamma[(n/2)+k-p-m]} \times r^{2p-2m}\tilde{f}_{k-2p}(\mathbf{x})$$

and then the canonical decomposition of $f_k(\mathbf{x})\in\mathscr{K}_k^{(n)}(\mathbf{x})$:

$$f_k(\mathbf{x}) = \sum_{p=0}^{[k/2]} r^{2p}\tilde{f}_{k-2p}(\mathbf{x})$$

with

$$\tilde{f}_{k-2p}(x) = \sum_{t=p}^{[k/2]}\frac{(-1)^{t-p}}{2^{2t}} \times\frac{[(n-2)/2+k-2p]\Gamma[(n-2)/2+k-p-t]}{\Gamma(p+1)\Gamma(t+1-p)\Gamma(n/2+k-p)} \times r^{2t-2p}\Delta^t[f_k(\mathbf{x})]. \tag{61}$$

This last relation gives the harmonic component $\tilde{f}_k(\mathbf{x})$ of $f_k(\mathbf{x})$ in the following form:

$$\tilde{f}_k(\mathbf{x}) = \sum_{p=0}^{[k/2]}\frac{(-1)^p}{2^{2p}p!}\frac{\Gamma[(n-2)/2+k-p]}{\Gamma[(n-2)/2+k]} \times r^{2p}\Delta^p[f_k(\mathbf{x})]. \tag{62}$$

In particular, the harmonic projection of x^k is

$$(\tilde{x}^k) = \sum_{p=0}^{[k/2]} \frac{(-1)^p}{2^{2p}p!} \frac{\Gamma(k+1)}{\Gamma(k+1-2p)} \times \frac{\Gamma[(n-2)/2+k+p]}{\Gamma[(n-2)/2+k]} r^{2p} x^{k-2p}. \quad (63)$$

which is nothing more than $z_k^{(n)}(x) = \mathscr{F}_k^{1,n-1}(x)$ on the isometric sphere ($r=1$), polynomial solution of Eq. (28).

2. Application of symmetric tensors

Let us then use the isomorphism between spaces $\mathscr{K}_k^{(n)}(\mathbf{x})$ and $SE_n^{\otimes k}$: If $f_k(\mathbf{x}) \in \mathscr{K}_k^{(n)}(\mathbf{x})$, let $f^{\alpha_1\alpha_2\cdots\alpha_k}$ be the corresponding tensor. With the Laplace–Beltrami operator, relation (61) is transposed immediately to the tensorial case. In particular, the harmonic projection of $f^{\alpha_1\alpha_2\cdots\alpha_k}$ is given by

$$\tilde{f}^{\alpha_1\alpha_2\cdots\alpha_k} = \sum_{p=0}^{[k/2]} \frac{(-1)^p}{2^{2p}p!} \frac{\Gamma[(n-2)/2+k-p]}{\Gamma[(n-2)/2+k]} \times r^{2p}\Delta^p(f^{\alpha_1\alpha_2\cdots\alpha_k}). \quad (64)$$

Using the Pichon–Marle tensor (56), the harmonic projection of the tensor $x^{\alpha_1\alpha_2\cdots\alpha_k} = x^{\alpha_1}x^{\alpha_2}\cdots x^{\alpha_k}$ is

$$\tilde{X}^{\alpha_1\alpha_2\cdots\alpha_k} = \sum_{p=0}^{[k/2]} \frac{(-1)^p}{2^p} \frac{\Gamma[(n-2)/2+k-p]}{\Gamma[(n-2)/2+k]} \times r^{2p}[X_{k-2p}, pg]^{\alpha_1\alpha_2\cdots\alpha_k}. \quad (65)$$

These results are applicable on a space endowed with a metric g of any type.

Relations (64) and (65) are, respectively, the covariant expressions of (62) and (63).

D. Application to classical polynomials (Ref. 2)

1. Covariant expression of zonal spherical functions

On the isotropic sphere ($r=1$), relation (65) leads immediately to the covariant expression of zonal spherical functions $z_k^{(n)}(x)$:

$$Z_{(n)}^{\alpha_1\alpha_2\cdots\alpha_k} = \sum_{p=0}^{[k/2]} \frac{(-1)^p}{2^p} \frac{\Gamma[(n-2)/2+k-p]}{\Gamma[(n-2)/2+k]} \times [X_{k-2p}, pg]^{\alpha_1\alpha_2\cdots\alpha_k}, \quad (66)$$

whose value on an axis $\mathbf{e}_\alpha$ ($\alpha = \alpha_1 = \alpha_2 = \cdots = \alpha_k$) is nothing more than $z_k^{(n)}(x)$ up to the sign due to the signature of the $g_{\alpha\beta}$ metric.

$$\mathscr{F}_k(x) = \sum_{p=0}^{[k/2]} F_{k-2p} x^{k-2p}.$$

There exists a symmetric tensor

$$\mathscr{F}^{\alpha_1\alpha_2\cdots\alpha_k} = \sum_{p=0}^{[k/2]} \frac{2^p p!\Gamma(k+1-2p)}{\Gamma(k+1)} \times [X_{k-2p}, pg]^{\alpha_1\alpha_2\cdots\alpha_k} \quad (67)$$

that we can understand as the covariant expression of $\mathscr{F}_k(x)$. Afterwards, we will assume that the polynomial $\mathscr{F}_k(x)$ is the projection of $\mathscr{F}^{\alpha_1\alpha_2\cdots\alpha_k}$ on $\mathbf{e}_\alpha$ such that $g_{\alpha\alpha} = +1$.

Then the transformation leading from the polynomial $\mathscr{F}_k(x)$ to its covariant form consists only in changing in $\mathscr{F}_k(x)$ each x term by a symmetric tensor of the same degree, multiplying by the tensor metric g as many times as necessary to obtain a tensor of order k, and in symmetrizing the while by dividing by the number of terms obtained.

Example: Let $\mathscr{F}_2(x) = 1 - x^2$. Its covariant expression is

$$\mathscr{F}^{\alpha\beta} = g^{\alpha\beta} - x^\alpha x^\beta. \quad (68)$$

In particular, in the Minkowski hyperboloïd case, (68) is the space projector.

We will call τ this transformation. We then have

$$\tau(X^{k-2p}) = \frac{2^p p!\Gamma(k+1-2p)}{\Gamma(k+1)} [X_{k-2p}, pg]^{\alpha_1\alpha_2\cdots\alpha_k}. \quad (69)$$

The τ transformation is obviously linear: If

$$\mathscr{F}_{k-2p}(x) = \sum_{t=0}^{[k-2p/2]} \mathscr{F}_{k-2p-2t} x^{k-2p-2t},$$

we will call $[\mathscr{F}_{k-2p}, pg]^{\alpha_1\alpha_2\cdots\alpha_k}$ the tensor obtained by a symmetrized product of tensor $\mathscr{F}^{\alpha_1\alpha_2\cdots\alpha_{k-2p}}$ [which is the covariant expression of $\mathscr{F}_{k-2p}(x)$] and p time the metric tensor g.

2. Covariant expression of fundamental equation

With projector (68), and Euler operators (51), the fundamental equation (28) of zonal functions may be written

$$[(g^{\alpha\beta} - x^\alpha x^\beta)\partial_{\alpha\beta} - (n-1)x^\alpha\partial_\alpha + k(k+n-2)] \times Z_{(n)}^{\alpha_1\alpha_2\cdots\alpha_k} = 0 \quad (70)$$

or

$$[\Delta - D_2 - (n-1)D_1 + k(k+n-2)] Z_{(n)}^{\alpha_1\alpha_2\cdots\alpha_k} = 0. \quad (71)$$

The operator

$$L = \Delta - D_2 - (n-1)D_1 + k(k+n-2) \quad (72)$$

is an invariant differential operator on the homogeneous space $\mathscr{M} = G/H$ (Ref. 6).

3. Application to classical polynomials

As the classical polynomials from Sec. I are proportional to the zonal spherical functions $z_{(n)}^{\alpha_i}$, we immediately obtain their covariant expression with the help of the τ transformation.

Hermite polynomials:

$$H^{\alpha_1\alpha_2\cdots\alpha_k} = \sum_{p=0}^{[k/2]} (-1)^p [X_{k-2p}, pg]^{\alpha_1\alpha_2\cdots\alpha_k}. \quad (73)$$

Gegembauer polynomials:

$$\frac{C_{n-2}^{\alpha_1\alpha_2\cdots\alpha_k}}{2} = \frac{1}{\Gamma(k+1)\Gamma[(n-2)/2]} \times \sum_{p=0}^{[k/2]} (-1)^p 2^{k-p} \Gamma\left(\frac{n-2}{2} + k - p\right) \times [X_{k-2p}, pg]^{\alpha_1\alpha_2\cdots\alpha_k}. \quad (74)$$

Tchebycheff polynomials:

$$T^{\alpha_1\alpha_2\cdots\alpha_k}=\frac{1}{\Gamma(k)}\sum_{p=0}^{[k/2]}(-1)^p 2^{k-1-p}\times\Gamma(k-p)\,[X_{k-2p},pg]^{\alpha_1\alpha_2\cdots\alpha_k}\quad(k\neq 0). \tag{75}$$

In a general way, all polynomials from Sec. I (particularly those being expressed with the hypergeometric function) can be written in a covariant form.

IV. PROPERTIES OF CLASSICAL POLYNOMIALS

A. Scalar case

1. $\mathscr{F}_k^{u,v}(x)$ polynomials

$$\mathscr{F}_k^{u,v}(x)=\sum_{p=0}^{[k/2]}\frac{(-1)^p}{2^{2p}p!u^p}\frac{\Gamma(k+1)}{\Gamma(k+1-2p)}\times\frac{\Gamma[(v-u)/2u+k-p]}{\Gamma[(v-u)/2u+k]}x^{k-2p}$$

is a particular solution of

$$(1-ux^2)\frac{d^2y}{dx^2}-vx\frac{dy}{dx}+k\,[v+(k-1)u]y=0.$$

Recurrence formulas:

$$(k+1)[v+(k-1)u]\mathscr{F}_k^{u,v}(x)=[v+(2k-1)u]\times[v+(2k+1)u]\,[x\mathscr{F}_{k+1}^{u,v}(x)-\mathscr{F}_{k+2}^{u,v}(x)].$$

$$(k+1)(1-ux^2)\mathscr{F}_k^{u,v+4u}(x)-vx\mathscr{F}_{k+1}^{u,v+2u}(x)+[v+(k+1)u]\mathscr{F}_{k+2}^{u,v}(x)=0.$$

Derivation formula:

$$\frac{d^p\mathscr{F}_k^{u,v}(x)}{dx^p}=\frac{\Gamma(k+1)}{\Gamma(k+1-p)}\mathscr{F}_{k-p}^{u,v+2pu}(x).$$

Transformation formula:

Any polynomials

$$\mathscr{F}_k(x)=\sum_{p=0}^{[k/2]}F_{k-2p}x^{k-2p}$$

can be expressed with the $[\mathscr{F}_{k-2p}^{u,v}(x)]_{0<p<[k/2]}$ by

$$\mathscr{F}_k(x)=\sum_{p=0}^{[k/2]}F_{k-2p}x^{k-2p}=\sum_{p=0}^{[k/2]}A_{k-2p}\mathscr{F}_{k-2p}^{u,v}(x),$$

with

$$A_{k-2p}=\sum_{t=0}^{P}\frac{1}{2^{2t}t!u^t}\frac{\Gamma(k+1-2p+2t)}{\Gamma(k+1-2p)}\times\frac{\Gamma[(v+u)/2u+k-2p]}{\Gamma[(v+u)/2u+k-2p+t]}F_{k-2p+2t},$$

$$F_{k-2p}=\sum_{t=0}^{P}\frac{(-1)^t}{2^{2t}t!u^t}\frac{\Gamma(k+1-2p+2t)}{\Gamma(k+1-2p)}\times\frac{\Gamma[(v-u)/2u+k-2p+t]}{\Gamma[(v-u)/2u+k-2p+2t]}A_{k-2p+2t}.$$

Particularly,

$$\mathscr{F}_k^{a,b}(x)=\sum_{p=0}^{[k/2]}B_{k-2p}\mathscr{F}_{k-2p}^{u,v}(x)$$

with

$$B_{k-2p}=\frac{1}{2^{2p}}\frac{\Gamma(k+1)}{\Gamma(k+1-2p)}\sum_{t=0}^{P}\frac{(-1)^t}{t!(p-t)!\,a^t u^{p-t}}\times\frac{\Gamma[(b-a)/2a+k-t]\Gamma[(v+u)/2u+k-2p]}{\Gamma[(b-a)/2a+k]\Gamma[(v+u)/2u+k-p-t]}$$

and (Nielsen formulas):

$$x^k=\sum_{p=0}^{[k/2]}\frac{1}{2^{2p}p!u^p}\frac{\Gamma(k+1)}{\Gamma(k+1-2p)}\times\frac{\Gamma[(v+u)/2u+k-2p]}{\Gamma[(v+u)/2u+k-p]}\mathscr{F}_{k-2p}^{u,v}(x).$$

Multiplication formula:

$$\mathscr{F}_{k_1}^{u_1,v_1}(x)\mathscr{F}_{k_2}^{u_2,v_2}(x)=\sum_{p=0}^{[k/2]}G_{k-2p}\mathscr{F}_{k-2p}^{a,b}(x)$$

with $k=k_1+k_2$, and

$$G_{k-2p}=\frac{1}{2^{2p}}\sum_{t=0}^{p}\left\{\frac{(-1)^t}{(p-t)!a^{p-t}}\frac{\Gamma(k+1-2t)}{\Gamma(k+1-2p)}\frac{\Gamma[(a+b)/2a+k-2p]}{\Gamma[(a+b)/2a+k-p-t]}\sum_{m_1+m_2=t}\frac{1}{m_1!m_2!u_1^{m_1}u_2^{m_2}}\times\frac{\Gamma(k_1+1)}{\Gamma(k_1+1-2m_1)}\frac{\Gamma(k_2+1)}{\Gamma(k_2+1-2m_2)}\frac{\Gamma[(v_1-u_1)/2u_1+k_1-m_1]\Gamma[(v_2-u_2)/2u_2+k_2-m_2]}{\Gamma[(v_1+u_1)/2u_1+k_1]\Gamma[(v_2-u_2)/2u_2+k_2]}\right\},$$

which gives for polynomials of the same kind (u,v):

$$\mathscr{F}_{k_1}^{u,v}(x)\mathscr{F}_{k_2}^{u,v}(x)=\sum_{p=0}^{[k/2]}G_{k-2p}\mathscr{F}_{k-2p}^{u,v}(x)$$

with $k=k_1+k_2$, and

$$G_{k-2p}=\frac{1}{2^{2p}u^p}\sum_{t=0}^{p}\left\{\frac{(-1)^t}{(p-t)!}\frac{\Gamma(k+1-2t)}{\Gamma(k+1-2p)}\frac{\Gamma[(v+u)/2u+k-2p]}{\Gamma[(v+u)/2u+k-p-t]}\sum_{m_1+m_2=t}\frac{1}{m_1!m_2!}\frac{\Gamma(k_1+1)}{\Gamma(k_1+1-2m_1)}\times\frac{\Gamma(k_2+1)}{\Gamma(k_2+1-2m_2)}\frac{\Gamma[(v-u)/2u+k_1-m_1]}{\Gamma[(v-u)/2u+k_1]}\frac{\Gamma[(v-u)/2u+k_2-m_2]}{\Gamma[(v-u)/2u+k_2]}\right\}.$$

2. Hermite polynomials

$$H_k(x) = \mathscr{F}_k^{0,1}(x) = \sum_{p=0}^{[k/2]} \frac{(-1)^p}{2^p p!} \frac{\Gamma(k+1)}{\Gamma(k+1-2p)} x^{k-2p}$$

is a particular solution of

$$\frac{d^2y}{dx^2} - x\frac{dy}{dx} + ky = 0.$$

Recurrence formula:

$$H_{k+2}(x) - xH_{k+1}(x) + (k+1)H_k(x) = 0.$$

Derivation formula:

$$\frac{d^p H_k(x)}{dx^p} = \frac{\Gamma(k+1)}{\Gamma(k+1-p)} H_{k-p}(x).$$

Transformation formula:

$$\mathscr{F}_k(x) = \sum_{p=0}^{[k/2]} F_{k-2p}x^{k-2p} = \sum_{p=0}^{[k/2]} A_{k-2p}H_{k-2p}(x),$$

with

$$A_{k-2p} = \sum_{t=0}^{p} \frac{1}{2^t t!} \frac{\Gamma(k+1-2p+2t)}{\Gamma(k+1-2p)} F_{k-2p+2t},$$

$$F_{k-2p} = \sum_{t=0}^{p} \frac{(-1)^t}{2^t t!} \frac{\Gamma(k+1-2p+2t)}{\Gamma(k+1-2p)} A_{k-2p+2t}.$$

Particularly,

$$x^k = \sum_{p=0}^{[k/2]} \frac{1}{2^p p!} \frac{\Gamma(k+1)}{\Gamma(k+1-2p)} H_{k-2p}(x).$$

Multiplication formula:

$$H_{k_1}(x)H_{k_2}(x) = \sum_{p=0}^{[k/2]} G_{k-2p}H_{k-2p}(x),$$

with $k = k_1 + k_2$, and

$$G_{k-2p} = \frac{1}{2^p} \sum_{t=0}^{p} \left\{ \frac{(-1)^t}{(p-t)!} \frac{\Gamma(k+1-2t)}{\Gamma(k+1-2p)} \times \sum_{m_1+m_2=t} \frac{1}{m_1!m_2!} \times \frac{\Gamma(k_1+1)\Gamma(k_2+1)}{\Gamma(k_1+1-2m_1)\Gamma(k_2+1-2m_2)} \right\}.$$

3. Gegembauer polynomials

$$C_k^l(x) = \frac{2^k\Gamma(k+l)}{\Gamma(k+1)\Gamma(p)} \mathscr{F}_k^{12l+1}(x) = \frac{1}{\Gamma(l)} \sum_{p=0}^{[k/2]} \frac{(-1)^p 2^{k-2p}}{p!} \frac{\Gamma(k+l-p)}{\Gamma(k+1-2p)} x^{k-2p},$$

solution of

$$(1-x^2)\frac{d^2y}{dx^2} - (2l+1)x\frac{dy}{dx} + k(k+2l)y = 0.$$

Recurrence formulas:

$$(k+2)C_{k+2}^l(x) = 2(k+l+1)xC_{k+1}^l(x) - (k+2l)C_k^l(x),$$

$$4l(l+1)(1-x^2)C_k^{l+2}(x) - 2l(l+1)xC_{k+1}^{l+1}(x) + (k+2l+2)(k+2)C_{k+2}^l(x) = 0.$$

Transformation formulas:

$$\mathscr{F}_k(x) = \sum_{p=0}^{[k/2]} F_{k-2p}x^{k-2p} = \sum_{p=0}^{[k/2]} A_{k-2p}C_{k-2p}^l(x)$$

with

$$A_{k-2p} = \frac{(k+l-2p)\Gamma(l)}{2^{k-2p}} \times \sum_{t=0}^{p} \frac{1}{2^{2t}t!} \frac{\Gamma(k+1-2p+2t)}{\Gamma(k+1+l-2p+t)} F_{k-2p+2t},$$

$$F_{k-2p} = \frac{2^{k-2p}}{\Gamma(l)\Gamma(k+1-2p)} \times \sum_{t=0}^{p} \frac{(-1)^t}{t!} \Gamma(k+l-2p+t)A_{k-2p+2t}.$$

Particularly,

$$x^k = \frac{\Gamma(k+1)\Gamma(l)}{2^k} \sum_{p=0}^{[k/2]} \frac{1}{p!} \times \frac{(k+l-2p)}{\Gamma(k+l+1-p)} C_{k-2p}^l(x).$$

Multiplication formula:

$$C_{k_1}^l(x)C_{k_2}^l(x) = \sum_{p=0}^{[k/2]} G_{k-2p}C_{k-2p}^l(x)$$

with $k = k_1 + k_2$, and

$$G_{k-2p} = \frac{l+k-2p}{\Gamma(l)} \sum_{t=0}^{p} \left\{ \frac{(-1)^t}{(p-t)!} \frac{\Gamma(k+1-2t)}{\Gamma(k+1+l-p-t)} \times \sum_{m_1+m_2=t} \frac{1}{m_1!m_2!} \times \frac{\Gamma(k_1+l-m_1)\Gamma(k_2+l-m_2)}{\Gamma(k_1+1-2m_1)\Gamma(k_2+1-2m_2)} \right\}.$$

4. Legendre polynomials

Legendre polynomials $P_k(x)$ are Gegembauer polynomials $C_k^l(x)$ with $l = 1/2$:

$$P_k(x) = C_k^{1/2}(x).$$

5. Tchebycheff polynomials

$$T_k(x) = 2^{k-1}\mathscr{F}_k^{1,1}(x) = k \sum_{p=0}^{[k/2]} \frac{(-1)^p 2^{k-1-2p}}{p!} \times \frac{\Gamma(k-p)}{\Gamma(k+1-2p)} x^{k-2p} \quad (k \neq 0),$$

solution of

$$(1-x^2)\frac{d^2y}{dx^2} - x\frac{dy}{dx} + k^2y = 0.$$

Recurrence formula:

$$T_{k+2}(x) - 2xT_{k+1}(x) + T_k(x) = 0.$$

Transformation formulas:

$$\mathscr{F}_k(x) = \sum_{p=0}^{[k/2]} F_{k-2p}x^{k-2p} = \sum_{p=0}^{[k/2]} A_{k-2p}T_{k-2p}(x),$$

with

$$A_{k-2p}=\frac{1}{2^{k-1-2p}}\sum_{t=0}^{p}\frac{1}{2^{2t}t!}\times\frac{\Gamma(k+1-2p+2t)}{\Gamma(k+1-2p+t)}F_{k-2p+2t},$$

$$F_{k-2p}=\frac{2^{k-1-2p}}{\Gamma(k+1-2p)}\sum_{t=0}^{p}\frac{(-1)^t}{t!}\times(k-2p+2t)\Gamma(k-2p+t)A_{k-2p+2t}.$$

Particularly,

$$x^k=\frac{1}{2^{k-1}}\sum_{p=0}^{[k/2]}\frac{1}{p!}\frac{\Gamma(k+1)}{\Gamma(k+1-p)}T_{k-2p}(x)$$

(with $T_0=\frac{1}{2}$).

Multiplication formula:

$$T_{k_1}(x)T_{k_2}(x)=\sum_{p=0}^{[k/2]}G_{k-2p}T_{k-2p}(x),$$

with $k=k_1+k_2$, and

$$G_{k-2p}=\frac{k_1k_2}{2}\sum_{t=0}^{p}\left\{\frac{(-1)^t}{(p-t)!}\frac{\Gamma(k+1-2t)}{\Gamma(k+1-p-t)}\times\sum_{m_1+m_2=t}\frac{1}{m_1!m_2!}\times\frac{\Gamma(k_1-m_1)\Gamma(k_2-m_2)}{\Gamma(k_1+1-2m_1)\Gamma(k_2+1-2m_2)}\right\}.$$

B. Covariant expressions

With the help of the τ transformation defined in Sec. III, we immediately obtain the covariant expressions of previous formulas. As τ is a linear transformation, coefficients A, B, and G appearing, respectively, in transformation and multiplication formulas are the same as in the scalar case.

1. $\mathcal{F}_k^{u,v}(x)$ polynomials

$$\mathcal{F}_{(u,v)}^{\alpha_1\alpha_2\cdots\alpha_k}=\sum_{p=0}^{[k/2]}\frac{(-1)^p}{2^pu^p}\frac{\Gamma[(v-u)/2u+k-p]}{\Gamma[(v-u)/2u+k]}\times[X_{k-2p},pg]^{\alpha_1\alpha_2\cdots\alpha_k}$$

solution of

$$\{\Delta-uD_2-vD_1+k[v+(k-1)u]\}\mathcal{F}_{(u,v)}^{\alpha_1\alpha_2\cdots\alpha_k}=0.$$

Recurrence formula:

$$2[v+(k-1)u][\mathcal{F}_k^{u,v},g]^{\alpha_1\alpha_2\cdots\alpha_{k+2}}=[v+(2k-1)u][v+(2k+1)u]\times\{[\mathcal{F}_{k+1}^{u,v},x]^{\alpha_1\alpha_2\cdots\alpha_{k+2}}-(k+2)\mathcal{F}_{(u,v)}^{\alpha_1\alpha_2\cdots\alpha_{k+2}}\}.$$

Transformation formulas:

$$\mathcal{F}^{\alpha_1\alpha_2\cdots\alpha_k}=\sum_{p=0}^{[k/2]}F_{k-2p}\frac{2^pp!\Gamma(k+1-2p)}{\Gamma(k+1)}\times[x_{k-2p},pg]^{\alpha_1\alpha_2\cdots\alpha_k}=\sum_{p=0}^{[k/2]}A_{k-2p}\frac{2^pp!\Gamma(k+1-2p)}{\Gamma(k+1)}\times[\mathcal{F}_{k-2p}^{u,v},pg]^{\alpha_1\alpha_2\cdots\alpha_k}.$$

Particularly,

$$\mathcal{F}_{(a,b)}^{\alpha_1\alpha_2\cdots\alpha_k}=\sum_{p=0}^{[k/2]}B_{k-2p}\frac{2^pp!\Gamma(k+1-2p)}{\Gamma(k+1)}\times[\mathcal{F}_{k-2p}^{u,v},pg]^{\alpha_1\alpha_2\cdots\alpha_k}.$$

and

$$X^{\alpha_1\alpha_2\cdots\alpha_k}=\sum_{p=0}^{[k/2]}\frac{1}{2^pu^p}\frac{\Gamma[(v+u)/2u+k-2p]}{\Gamma[(v+u)/2u+k-p]}\times[\mathcal{F}_{k-2p}^{u,v},pg]^{\alpha_1\alpha_2\cdots\alpha_k}.$$

Multiplication formula:

$$(\mathcal{F}_{k_1}^{u_1,v_1}\mathcal{F}_{k_2}^{u_2,v_2})^{\alpha_1\alpha_2\cdots\alpha_k}=\sum_{p=0}^{[k/2]}G_{k-2p}\frac{2^pp!\Gamma(k+1-2p)}{\Gamma(k+1)}\times[\mathcal{F}_{k-2p}^{a,b},pg]^{\alpha_1\alpha_2\cdots\alpha_k},$$

where $k=k_1+k_2$.

2. Hermite polynomials

$$H^{\alpha_1\alpha_2\cdots\alpha_k}=\sum_{p=0}^{[k/2]}(-1)^p[X_{k-2p},pg]^{\alpha_1\alpha_2\cdots\alpha_k}$$

solution of

$$(\Delta-D_1+k)H^{\alpha_1\alpha_2\cdots\alpha_k}=0.$$

Recurrence formulas:

$$2[H_k,g]^{\alpha_1\alpha_2\cdots\alpha_{k+2}}=[H_{k+1},x]^{\alpha_1\alpha_2\cdots\alpha_{k+2}}-(k+2)H^{\alpha_1\alpha_2\cdots\alpha_{k+2}}.$$

Transformation formula:

$$\mathcal{F}^{\alpha_1\alpha_2\cdots\alpha_k}=\sum_{p=0}^{[k/2]}F_{k-2p}\frac{2^pp!\Gamma(k+1-2p)}{\Gamma(k+1)}\times[X_{k-2p},pg]^{\alpha_1\alpha_2\cdots\alpha_k}=\sum_{p=0}^{[k/2]}A_{k-2p}\frac{2^pp!\Gamma(k+1-2p)}{\Gamma(k+1)}\times[H_{k-2p},pg]^{\alpha_1\alpha_2\cdots\alpha_k}.$$

Particularly,

$$X^{\alpha_1\alpha_2\cdots\alpha_k}=\sum_{p=0}^{[k/2]}[H_{k-2p},pg]^{\alpha_1\alpha_2\cdots\alpha_k}.$$

Multiplication formula:

$$(H_{k_1}H_{k_2})^{\alpha_1\alpha_2\cdots\alpha_k}=\sum_{p=0}^{[k/2]}G_{k-2p}\frac{2^pp!\Gamma(k+1-2p)}{\Gamma(k+1)}\times[H_{k-2p},pg]^{\alpha_1\alpha_2\cdots\alpha_k},$$

where $k=k_1+k_2$.

3. Gegenbauer polynomials

$$C_l^{\alpha_1\alpha_2\cdots\alpha_k}=\frac{1}{\Gamma(k+1)\Gamma(l)}\sum_{p=0}^{[k/2]}(-1)^p\times2^{k-p}\Gamma(l+k-p)[X_{k-2p},pg]^{\alpha_1\alpha_2\cdots\alpha_k}$$

solution of

$$\{\Delta - D_2 - (2l+1)D_1 + k(k+2l)\}C_l^{\alpha_1\alpha_2\cdots\alpha_k} = 0.$$

Recurrence formula:

$$2(k+2l)[C_k^l, g]^{\alpha_1\alpha_2\cdots\alpha_{k+2}} = 2(k+l+1)(k+1)[C_{k+1}^l, X]^{\alpha_1\alpha_2\cdots\alpha_{k+2}} - (k+1)(k+2)^2 C_{(l)}^{\alpha_1\alpha_2\cdots\alpha_{k+2}}.$$

Transformation formula:

$$\mathcal{F}^{\alpha_1\alpha_2\cdots\alpha_k} = \sum_{p=0}^{[k/2]} F_{k-2p} \frac{2^p p!\Gamma(k+1-2p)}{\Gamma(k+1)} \times [X_{k-2p}, pg]^{\alpha_1\alpha_2\cdots\alpha_k} = \sum_{p=0}^{[k/2]} A_{k-2p} \frac{2^p p!\Gamma(k+1-2p)}{\Gamma(k+1)} \times [C_{k-2p}^l, pg]^{\alpha_1\alpha_2\cdots\alpha_k}.$$

Particularly,

$$X^{\alpha_1\alpha_2\cdots\alpha_k} = \Gamma(l) \sum_{p=0}^{[k/2]} \frac{(k+l-2p)\Gamma(k+1-2p)}{2^{k-p}\Gamma(k+l+1-p)} \times [C_{k-2p}^l, pg]^{\alpha_1\alpha_2\cdots\alpha_k}.$$

Multiplication formula:

$$(C_{k_1}^l C_{k_2}^l)^{\alpha_1\alpha_2\cdots\alpha_k} = \sum_{p=0}^{[k/2]} G_{k-2p} \frac{2^p p!\Gamma(k+1-2p)}{\Gamma(k+1)} \times [C_{k-2p}^l, pg]^{\alpha_1\alpha_2\cdots\alpha_k},$$

where $k = k_1 + k_2$.

4. Tchebycheff polynomials

$$T^{\alpha_1\alpha_2\cdots\alpha_k} = \frac{1}{\Gamma(k)} \sum_{p=0}^{[k/2]} (-1)^p 2^{k-1-p} \times \Gamma(k-p)[X_{k-p}, pg]^{\alpha_1\alpha_2\cdots\alpha_k},$$

where $k \neq 0$. This has a solution of

$$(\Delta - D_2 - D_1 + k^2)T^{\alpha_1\alpha_2\cdots\alpha_k} = 0.$$

Recurrence formula:

$$2[T_k, g]^{\alpha_1\alpha_2\cdots\alpha_{k+2}} = 2(k+1)[T_{k+1}, X]^{\alpha_1\alpha_2\cdots\alpha_{k+2}} - (k+1)(k+2)T^{\alpha_1\alpha_2\cdots\alpha_{k+2}}.$$

Transformation formula:

$$\mathcal{F}^{\alpha_1\alpha_2\cdots\alpha_k} = \sum_{p=0}^{[k/2]} F_{k-2p} \frac{2^p p!\Gamma(k+1-2p)}{\Gamma(k+1)} \times [X_{k-2p}, pg]^{\alpha_1\alpha_2\cdots\alpha_k} = \sum_{p=0}^{[k/2]} A_{k-2p} \frac{2^p p!\Gamma(k+1-2p)}{\Gamma(k+1)} \times [T_{k-2p}, pg]^{\alpha_1\alpha_2\cdots\alpha_k}.$$

Particularly,

$$X^{\alpha_1\alpha_2\cdots\alpha_k} = \frac{1}{2^{k-1}} \sum_{p=0}^{[k/2]} \frac{2^p\Gamma(k+1-2p)}{\Gamma(k+1)} \times [T_{k-2p}, pg]^{\alpha_1\alpha_2\cdots\alpha_k}.$$

Multiplication formula:

$$(T_{k_1} T_{k_2})^{\alpha_1\alpha_2\cdots\alpha_k} = \sum_{p=0}^{[k/2]} G_{k-2p} \frac{2^p p!\Gamma(k+1-2p)}{\Gamma(k+1)} \times [T_{k-2p}, pg]^{\alpha_1\alpha_2\cdots\alpha_k},$$

where $k = k_1 + k_2$.

ACKNOWLEDGMENTS

I am pleased to express my deepest gratitude to Professor Lichnerowicz and Professor Lutz for all their advice and constant support.

[1] *Séminaire d'Analyse Harmonique de Nancy-Strasbourg* (Springer-Verlag, Berlin, 1975).

[2] J. C. Lucquiaud, C. R. Acad. Sci. Paris **287**, 67 (1978); **292**, 653, 855 (1981); **294**, 549 (1982); **299**, 467, (1984); J. Math. Pures Appl. **63**, 265 (1984).

[3] N. Ja Vilemkin, *Fonctions Spéciales et Théorie de la Représentation* (Dunod, Paris, 1969).

[4] J. Kampé de Feriet, *Fonctions de la Physique Mathématique* (CNRS, Paris, 1957).

[5] C. M. Malle, Ann. Inst. H. Poincaré, **1**, 67 (1969).

[6] A. Lichnerowicz, Ann. Sci. Ecole Norm. Sup. **81**, 341 (1964).

Complete integrability and analytic solutions of a KdV-type equation

Zhi-xiong Chen
Department of Applied Mathematics, Shanghai Jiao Tong University, People's Republic of China

Ben-yu Guo (Pen-yu Kuo)
Department of Mathematics, Shanghai University of Science and Technology, People's Republic of China

Long-wan Xiang
Department of Applied Mathematics, Shanghai Jiao Tong University, People's Republic of China

(Received 13 September 1989; accepted for publication 1 August 1990)

The complete integrability of the variable coefficient version of a KdV equation via the Painlevé approach is analyzed. Through the Painlevé–Bäcklund equations, its auto-Bäcklund transformation, Lax pairs, symmetry, strong symmetry, bilinear form, and analytic solutions are obtained.

I. INTRODUCTION

The variable coefficient version of a KdV equation is

$$u_t + \alpha(t)uu_x + \beta(t)u_{xxx} = 0, \tag{1.1}$$

which describes the breaking phenomena of soliton-like waves in a varied depth shallow-water tunnel. Many authors have studied this equation analytically and numerically (see Refs. 1–6). Abellanas and Galindo[4] showed the existence of a change of variables, which transforms any nonautonomous KdV-type equations with local nontrivial conserved densities of arbitrary higher order into autonomous KdV-type equations. So, Eq. (1.1) may be solved from the solutions of a corresponding autonomous KdV-type equation. But as mentioned in Ref. 3, such change does not preserve the boundedness conditions. Therefore, certain effective methods fail in this case. Recently, some authors used the Painlevé approach for Eq. (1.1). By such analysis, Joshi[3] concluded that Eq. (1.1) has Painlevé property only when

$$\beta(t) = \alpha(t)\left[aD_t^{-1}\alpha(t) + b\right]. \tag{1.2}$$

The above constraint relation is the same as the result obtained by Grimshaw.[2] On the other hand, Nirmale *et al.*[6] considered the special case with $\alpha(t) = \alpha_0 t^{\gamma}$, $\beta(t) = \beta_0 t^{2\gamma+1}$. They showed its Painlevé property, auto-Bäcklund transformation, and Lax pairs, meanwhile they got some exact solutions by similarity transformation.

In this paper, we use the Painlevé approach[7] to analyze the complete integrability and analytic solutions. We obtain its auto-Bäcklund transformation, Lax-pairs, symmetry, strong symmetry, and bilinear form. All these results are delivered in Sec. II. In Sec. III, we follow the idea of Conte[8] to construct its analytic solutions by the Painlevé–Bäcklund equations.

II. COMPLETE INTEGRABILITY

We say that Eq. (1.1) possesses the Painlevé property, if the following Lorentz series expansion

$$u = \sum_{j=0}^{\infty} u_j \phi^{j-p} \tag{2.1}$$

is single valued about the movable solution singular manifold $\phi = 0$, that is, p is a positive integer, ϕ is analytic and noncharacteristic ($\phi_x\phi_t \neq 0$); all recursion relations for u_j are self-consistent and there are enough free functions in (2.1) in the sense of Cauchy–Kowalevskia theorem.[7]

By substituting (2.1) into (1.1) and analyzing leading parts, we find $p = 2$, and the following recursion relation:

$$\begin{aligned} &u_{j-3,t} + (j-4)u_{j-2}\phi_t + \alpha(t)\left(\sum_{k=0}^{j} u_{j-k}\left[u_{k-1,x} + (k-2)(u_k\phi_x)\right]\right) + \beta(t)\left[u_{j-3,xxx} + 3(j-4)u_{j-2,xx}\phi_x\right. \\ &\quad + 3(j-4)u_{j-2,x}\phi_{xx} + (j-4)u_{j-2}\phi_{xxx} + 3(j-4)(j-3)u_{j-1,x}\phi_x^2 + 3(j-4)(j-3)u_{j-1}\phi_x\phi_{xx} \\ &\quad \left. + (j-4)(j-3)(j-2)u_j\phi_x^3\right] = 0. \end{aligned} \tag{2.2}$$

Clearly, we have for $j = 0$,

$$u_0 = -\delta(t)\phi_x^2, \tag{2.3}$$

with

$$\delta(t) = 12\beta(t)/\alpha(t) = 12\left[aD_t^{-1}\alpha(t) + b\right],$$

and thus (2.2) becomes:

 0022-2488/90/122851-05$03.00

$$(j+1)(j-4)(j-6)\beta(t)\phi_x^3 u_j = -u_{j-3,t} - (j-4)u_{j-2}\phi_t - \alpha(t)\left(\sum_{k=1}^{j-1} u_{j-k}(u_{k-1,x} + (k-2)u_k\phi_x) + u_0 u_{j-1,x}\right)$$
$$-\beta(t)\{u_{j-3,xxx} + 3(j-4)u_{j-2,xx}\phi_x + 3(j-4)u_{j-2,x}\phi_{xx}$$
$$+ (j-4)u_{j-2}\phi_{xxx} + 3(j-4)(j-3)u_{j-1,x}\phi_x^2 + 3(j-4)(j-3)u_{j-1}\phi_{xx}\phi_x\} = 0. \tag{2.4}$$

The resonance points are -1, 4, and 6. The point $j=-1$ corresponds to the arbitrary singular function ϕ. Expression (2.4) holds automatically for $j=4$. While the compatability for $j=6$ leads to constraint condition (1.2) that was given in Ref. 3 by reduction $\phi = x + \psi(t)$. But for delivering its complete integrability and constructing analytic solutions, we cannot use such reduction.

We now put $j=1,2,3$ in (2.4) and get

$$u_1 = \delta(t)\phi_{xx}, \tag{2.5}$$

$$\phi_x\phi_t + \alpha(t)\phi_x^2 u_2 + \beta(t)(4\phi_x\phi_{xxx} - 3\phi_{xx}^2) = 0, \tag{2.6}$$

and

$$\phi_{xt} + \alpha(t)\phi_{xx}u_2 + \beta(t)\phi_{xxxx} + \frac{\delta'(t)}{\delta(t)}\phi_x - \alpha(t)\phi_x^2 u_3 = 0. \tag{2.7}$$

Proposition 1: Equation (1.1) has the auto-Bäcklund transformation in the form:

$$u = 12[aD_t^{-1}\alpha(t) + b]D_x^2(\ln\phi) + u_2, \tag{2.8}$$

where u and u_2 satisfy Eq. (1.1), and ϕ satisfies

$$\phi_x\phi_t + \beta(t)(4\phi_x\phi_{xxx} - 3\phi_{xx}^2) + \alpha(t)\phi_x^2 u_2 = 0, \tag{2.9}$$

$$\phi_{xt} + \beta(t)\phi_{xxxx} + \frac{\alpha(t)}{D_t^{-1}\alpha(t)}\phi_x + \alpha(t)\phi_{xx}u_2 = 0. \tag{2.10}$$

Proof: Set arbitrary functions $u_4 = u_6 = 0$ and demand $u_3 = 0$. Then we have from (2.4) that $u_j = 0$, $\forall j \geqslant 3$, and thus

$$u = u_0/\phi^2 + u_1/\phi + u_2 = \delta(t)D_x^2(\ln\phi) + u_2,$$

which is just (2.8). Next by letting $u_3 = 0$ in (2.7), we obtain (2.10). ■

One may doubt about the compatability of Eqs. (2.9), (2.10), and (1.1). Indeed the former two lead to the third. This will be explained in the following proposition.

Proposition 2: Equation (1.1) has the Lax pairs

$$\psi_{xx} = -\frac{1}{6\beta(t)}\left[\alpha(t)u - \frac{\delta'(t)}{\delta(t)}x + \lambda(t)\right]\psi, \tag{2.11}$$

$$\psi_t = -\frac{1}{3}\left[\alpha(t)u + 2\frac{\delta'(t)}{\delta(t)}x - 2\lambda(t)\right]\psi_x + \frac{1}{6}\left[\alpha(t)u_x + 2\frac{\delta'(t)}{\delta(t)}\right]\psi, \tag{2.12}$$

where $\lambda(t)$ is the spectral function of (1.1).

Proof: We shall use the "Schwarzian derivative-scattering" method. Eliminating u_2 in Eqs. (2.9) and (2.10), we obtain

$$D_x\left[\frac{\phi_t}{\phi_x} + \beta(t)\{\phi:x\} + \frac{\delta'(t)}{\delta(t)}x\right] = 0,$$

where

$$\{\phi:x\} = D_x\left(\frac{\phi_{xx}}{\phi_x}\right) - \frac{1}{2}\left(\frac{\phi_{xx}}{\phi_x}\right)^2, \tag{2.13}$$

which is called Schwarzian derivative. Thus

$$\frac{\phi_t}{\phi_x} + \beta(t)\{\phi:x\} + \frac{\delta'(t)}{\delta(t)}x = \lambda(t). \tag{2.14}$$

Now let

$$\phi = \psi_1/\psi_2, \tag{2.15}$$

where we ask ψ_1 and ψ_2 to satisfy the same second scattering problem:

$$\psi_{xx} = U\psi, \tag{2.16}$$

$$\psi_t = V\psi_x + W\psi. \tag{2.17}$$

Then (2.14) becomes

$$V - 2\beta(t)U + \frac{\delta'(t)}{\delta(t)}x = \lambda(t). \tag{2.18}$$

On the other hand, the compatability of (2.16) and (2.17) shows

$$V_t = U_x V + 2UV_x + W_{xx}, \tag{2.19}$$

$$W = -V_x/2. \tag{2.20}$$

Therefore,

$$U_t = 6\beta(t)UU_x - \beta(t)U_{xxx} - \frac{\delta'(t)}{\delta(t)}xU_x + \lambda(t)U_x - 2\frac{\delta'(t)}{\delta(t)}U. \tag{2.21}$$

Furthermore, let

$$U = -[\alpha(t)/6\beta(t)]u + Y(x,t), \tag{2.22}$$

where u satisifies Eq. (1.1). Then

$$Y = \frac{1}{6\beta(t)}\left[\frac{\delta'(t)}{\delta(t)}x - \lambda(t)\right] \tag{2.23}$$

and thus

$$U = -\frac{1}{6\beta(t)}\left[\alpha(t)u - \frac{\delta'(t)}{\delta(t)}x + \lambda(t)\right]. \tag{2.24}$$

So from (2.19) and (2.20), we have

$$V = -\frac{1}{3}\left[\alpha(t)u + 2\frac{\delta'(t)}{\delta(t)}x - 2\lambda(t)\right] \tag{2.25}$$

and

$$W = \frac{1}{6}\left[\alpha(t)u_x + 2\frac{\delta'(t)}{\delta(t)}x\right]. \tag{2.26}$$

Hence, the Lax pairs (2.11) and (2.12) are obtained. ■

Proposition 3: Equation (1.1) has the symmetry

$$u_1 = \delta(t)\phi_{xx} \tag{2.27}$$

and strong symmetry operator

$$R = \beta(t)D_x^2 + \frac{2}{3}\left[\alpha(t)u - \frac{\delta'(t)}{\delta(t)}x\right] - \frac{1}{3}\left[\alpha(t)u_x - \frac{\delta'(t)}{\delta(t)}\right]D_x^{-1}. \tag{2.28}$$

In addition, if we take $K_0 = u_0$ to be the first symmetry, then we have a family of symmetries:

$$K_n = R^n[K_0], \quad n = 0,1,2,\ldots. \tag{2.29}$$

Proof: The idea of the proof is due to Strampp.[9] By substituting

$$u = u_0/\phi^2 + u_1/\phi + u_2$$

into (1.1) and analyzing the coefficient of the term ϕ^{-1}, we discover that u_1 satisfies the linearized equation of (1.1):

$$T_t + [\alpha(t)(uD_x + u_x) + \beta(t)D_x^3]T = 0, \tag{2.30}$$

which indicates that $u_1 = \delta(t)\phi_{xx}$ is a symmetry of (1.1). use of (2.5), (2.15), (2.11) and (2.12), (2.30) turns to another form

$$\left\{\beta(t)D_x^2 + \frac{2}{3}\left[\alpha(t)u - \frac{\delta'(t)}{\delta(t)}x\right] - \frac{1}{3}\left[\alpha(t)u_x - \frac{\delta'(t)}{\delta(t)}D_x^{-1}\right]\right\}T = H(\lambda)T, \tag{2.31}$$

which implies that the left hand is a strong symmetry operator. Then it is easy to get the conclusion (2.29). ■

Proposition 4: Equation (1.1) has the bilinear form:

$$(\tau_{xt}\tau - \tau_x\tau_t) + \beta(t)(\tau\tau_{xxxx} - 4\tau_{xxx}\tau_x + 3\tau_{xx}^2) + \frac{\delta'(t)}{\delta(t)}\tau\tau_x = 0. \tag{2.32}$$

In addition, if we set

$$u^{(n)} = \delta(t)D_x^2(\ln\tau_n) \tag{2.33}$$

and rewrite (2.8) as

$$u^{(n)} = \delta(t)D_x^2(\ln\phi_n) + u^{(n-1)}, \tag{2.34}$$

then

$$\tau_n = \tau_{n-1}\phi_n \tag{2.35}$$

and thus

$$\tau_n = \prod_{j=1}^{n}\phi_j. \tag{2.36}$$

Proof: It is easy to verify (2.32) by substituting (2.33) into (1.1). The rest of the proof is clear. ■

III. ANALYTIC SOLUTIONS

We have introduced the Schwarzian derivative

$$S \doteq \{\phi:x\} = \frac{\phi_{xxx}}{\phi_x} - \frac{3}{2}\left(\frac{\phi_{xx}}{\phi_x}\right)^2, \tag{3.1}$$

which is invariant under the homographic transformation. In this section, we give another homographic invariant

$$Z \doteq -\phi_t/\phi_x, \tag{3.2}$$

which is called velocity dimension. If we adapt "minus square transformation:"

$$\phi_x = V^{-2}, \tag{3.3}$$

then Eqs. (3.1) and (3.2) lead to

$$V_{xx} + (S/2)V = 0, \tag{3.4}$$

$$V_t + ZV_x - \tfrac{1}{2}Z_xV = 0. \tag{3.5}$$

The above system is linear for V when S and Z are known. By substituting (3.1) and (3.2) into (1.1), (2.9), and (2.10), we shall get the expressions of S and Z that are called Painlevé–Bäcklund equations. Next from such equations, we can deliver the explicit expression of V, ϕ, as well as u_2. For convenience, we define

$$Q = -\phi_{xx}/2\phi_x, \tag{3.6}$$

We get from (2.9) and (2.10) that

$$u_2 = [1/\alpha(t)][Z - 4\beta(t)S - 12\beta(t)Q^2] \tag{3.7}$$

and

$$Z_x = \beta(t)S_x + \frac{\delta'(t)}{\delta(t)}. \tag{3.8}$$

On the other hand, by substituting (3.7) into (1.1), we find that

$$Z_t - 4\tilde{S}_t - \frac{\alpha'(t)}{\alpha(t)}(Z - 4\tilde{S}) + (Z - 4\tilde{S})(Z_x - 4\tilde{S}_x) + \beta(t)Z_{xxx} - 4\beta(t)\tilde{S}_{xxx} - 18\tilde{S}_x\tilde{S} + Q[12\tilde{S}_{xx} - 12\beta(t)Z_{xx}] + 12\beta(t)Q^2\left[Z_x - \frac{\tilde{S}_x}{\beta(t)} - 1 + \frac{\alpha'(t)}{\alpha(t)}\right] = 0,$$

where $\tilde{S} = \beta(t)S$. We multiply (3.8) by $(QQ_x - Q^2)$ and put the result and the above equation together. Then

$$Z_t - 4\tilde{S}_t - \frac{\alpha'(t)}{\alpha(t)}(Z - 4\tilde{S}) + (Z - 4\tilde{S})(Z_x - 4\tilde{S}_x) + \beta(t)Z_{xxx} - 4\beta(t)\tilde{S}_{xxx} - 18\tilde{S}_x\tilde{S} = 0. \tag{3.9}$$

Furthermore, the compatability condition of (3.1) and (3.2) leads to

$$\tilde{S}_t + \beta(t)Z_{xxx} + 2Z_x\tilde{S} + Z\tilde{S}_x - [\beta'(t)/\beta(t)]\tilde{S} = 0. \tag{3.10}$$

We now try to solve Eqs. (3.8)–(3.10). We have from (3.8) that

$$Z = \tilde{S} + \frac{\delta'(t)}{\delta(t)}x + \gamma(t), \tag{3.11}$$

where $\gamma(t)$ is an arbitrary function. Thus (3.9) and (3.10) read

$$\tilde{S}_t + \beta(t)\tilde{S}_{xxx} + 3\tilde{S}_x\tilde{S} + \left(2\frac{\delta'(t)}{\delta(t)} - \frac{\beta'(t)}{\beta(t)}\right)\tilde{S} + \left(\frac{\delta'(t)}{\delta(t)}x + \gamma(t)\right)\tilde{S}_x = 0 \tag{3.12}$$

and

$$-3[\tilde{S}_t + \beta(t)\tilde{S}_{xx} + 3\tilde{S}_x\tilde{S}] - 3\left[2\frac{\delta'(t)}{\delta(t)} - \frac{\alpha'(t)}{\alpha(t)}\right]\tilde{S} - 3\left[\frac{\delta'(t)}{\delta(t)}x + \gamma(t)\right]\tilde{S}_x$$

$$+x\left[\left(\frac{\delta'(t)}{\delta(t)}\right)'+\frac{\delta'(t)}{\delta(t)}\left(\frac{\delta'(t)}{\delta(t)}-\frac{\alpha'(t)}{\alpha(t)}\right)\right]$$
$$+\gamma'(t)+\gamma(t)\left(\frac{\delta'(t)}{\delta(t)}-\frac{\alpha'(t)}{\alpha(t)}\right)=0, \tag{3.13}$$

which gives

$$\tilde{S}_t+\beta(t)\tilde{S}_{xxx}+3\tilde{S}_x\tilde{S}+\left(2\frac{\delta'(t)}{\delta(t)}-\frac{\beta'(t)}{\beta(t)}\right)\tilde{S}$$
$$+\left(\frac{\delta'(t)}{\delta(t)}x+\gamma(t)\right)\tilde{S}_x=0 \tag{3.14}$$

and

$$x\left[\left(\frac{\delta'(t)}{\delta(t)}\right)'+\frac{\delta'(t)}{\delta(t)}\left(\frac{\delta'(t)}{\delta(t)}-\frac{\alpha'(t)}{\alpha(t)}\right)\right]+\gamma'(t)$$
$$+\gamma(t)\left(\frac{\delta'(t)}{\delta(t)}-\frac{\alpha'(t)}{\alpha(t)}\right)=0. \tag{3.15}$$

Obviously, (3.15) is equivalent to

$$\left(\frac{\delta'(t)}{\delta(t)}\right)'+\frac{\delta'(t)}{\delta(t)}\left(\frac{\delta'(t)}{\delta(t)}-\frac{\alpha'(t)}{\alpha(t)}\right)=0, \tag{3.16}$$

$$\gamma'(t)+\gamma(t)\left(\frac{\delta'(t)}{\delta(t)}-\frac{\alpha'(t)}{\alpha(t)}\right)=0. \tag{3.17}$$

It is not easy to solve Eqs. (3.14), (3.16), and (3.17). We confine our discussion to some special cases.

Case I: $a=0$ in (1.2). In this case, $\delta(t)=b$. Then, (3.16) becomes an identity and (3.17) leads to

$$\gamma(t)=f(x)\alpha(t), \tag{3.18}$$

where $f(x)$ is an arbitrary function. Thus we have from (3.14) that

$$\tilde{S}_t+\beta(t)\tilde{S}_{xxx}+3\tilde{S}_x\tilde{S}-\frac{\alpha'(t)}{\alpha(t)}\tilde{S}+\gamma(t)\tilde{S}_x=0. \tag{3.19}$$

If we further let $\tilde{S}=\tilde{S}(t)$, then the above expression changes into

$$\tilde{S}_t-\frac{\alpha'(t)}{\alpha(t)}\tilde{S}=0,$$

which gives

$$\tilde{S}=g(x)\alpha(t), \tag{3.20}$$

where $g(x)$ is an arbitrary function. Thus

$$S=g(x)/b. \tag{3.21}$$

From (3.11),

$$Z=[f(x)+g(x)]\alpha(t). \tag{3.22}$$

If $g(x)/b$ and $f(x)+g(x)$ are constants, say $-k_0^2/2$ and c_0, respectively, then

$$S=-k_0^2/2, \quad Z=c_0\alpha(t).$$

Thus (3.4) and (3.5) become

$$V_{xx}-(k_0^2/4)V=0, \tag{3.23}$$

$$V_t+c_0\alpha(t)V_x=0. \tag{3.24}$$

Therefore,

$$V=Ae^{(k_0/2)\xi}+Be^{-(k_0/2)\xi}, \quad \xi=x-c_0D_t^{-1}\alpha(t)+d_0, \tag{3.25}$$

where A, B, and d_0 are constants. Then from (3.3),

$$\phi=\frac{Ce^{(k_0/2)\xi}+De^{-(k_0/2)\xi}}{Ae^{(k_0/2)\xi}+Be^{-(k_0/2)\xi}}, \tag{3.26}$$

where C and D are arbitrary constants provided that

$$CB-AD=1/k_0.$$

Hence, (3.7) leads to

$$u_2=-3bk_0^2\left[\frac{Ae^{(k_0/2)\xi}-Be^{-(k_0/2)\xi}}{Ae^{(k_0/2)\xi}+Be^{-(k_0/2)\xi}}\right]^2+2bk_0^2+c_0. \tag{3.27}$$

Using the auto-Bäcklund transformation (2.8), we find that the obtained solution u is the same as (3.27) provided that A and B are replaced by C and D.

Similarly, we can look for other solutions of $\tilde{S}$ and Z through (3.19) and (3.11), and get other analytic solutions.

Case II: $b=0$ in (1.2). In this case $\delta(t)=aD_t^{-1}\alpha(t)$. Then (3.16) becomes an identity, and (3.17) reads

$$\gamma'(t)-\gamma(t)\left(\frac{\delta'(t)}{\delta(t)}\right)'\frac{\delta(t)}{\delta'(t)}=0,$$

which leads to

$$\gamma(t)=f_1(x)\frac{\delta'(t)}{\delta(t)}, \tag{3.28}$$

where $f_1(x)$ is an arbitrary function. If we take $\tilde{S}=\tilde{S}(t)$, then (3.14) becomes

$$\tilde{S}_t+\tilde{S}\left[\frac{\delta'(t)}{\delta(t)}-\frac{\alpha'(t)}{\alpha(t)}\right]=0,$$

which is equivalent to

$$\tilde{S}_t-\tilde{S}\left(\frac{\delta'(t)}{\delta(t)}\right)'\left(\frac{\delta(t)}{\delta'(t)}\right)=0.$$

Hence,

$$\tilde{S}=g_1(x)\frac{\delta'(t)}{\delta(t)}, \tag{3.29}$$

where $g_1(x)$ is an arbitrary function. Thus

$$S=[g_1(x)/a][D_t^{-1}\alpha(t)]^2 \tag{3.30}$$

and then

$$Z=[f_1(x)+g_1(x)+x][\alpha(t)/D_t^{-1}\alpha(t)]. \tag{3.31}$$

If we set $f_1(x)+g_1(x)=0$, $g_1(x)/a=-k_1^2/2$ where k_1 is a constant, then

$$S=-[k_1/2D_t^{-1}\alpha(t)]^2, \quad Z=[\alpha(t)/D_t^{-1}\alpha(t)]x.$$

Therefore, (3.4) and (3.5) turn into

$$V_{xx}-[k_1/2D_t^{-1}\alpha(t)]^2V=0, \tag{3.32}$$

$$V_t+[\alpha(t)/D_t^{-1}\alpha(t)][xV_x-\tfrac{1}{2}V]=0. \tag{3.33}$$

Hence,

$$V=(A_1e^{k_1\eta/2D_t^{-1}\alpha(t)}+B_1e^{-k_1\eta/2D_t^{-1}\alpha(t)})(D_t^{-1}\alpha(t))^{1/2},$$
$$\eta=x, \tag{3.34}$$

where A_1 and B_1 are constants. Therefore

$$\phi=\frac{C_1e^{k_1\eta/2D_t^{-1}\alpha(t)}+D_1e^{-k_+\eta/2D_t^{-1}\alpha(t)}}{A_1e^{k_1\eta/2D_t^{-1}\alpha(t)}+B_1e^{-k_1\eta/2D_t^{-1}\alpha(t)}}, \tag{3.35}$$

where C_1 and D_1 are arbitrary constants provided that

$$C_1B_1 - A_1D_1 = D_t^{-1}\alpha(t)/k_1.$$

Finally,

$$u_2 = \frac{1}{D_t^{-1}\alpha(t)}\left[x + 2ak_1^2 - 3ak_1^2 \times\left(\frac{A_1e^{k_1\eta/2D_t^{-1}\alpha(t)} - B_1e^{-k_1\eta/2D_t^{-1}\alpha(t)}}{A_1e^{k_1\eta/2D_t^{-1}\alpha(t)} - B_1e^{-k_1\eta/2D_t^{-1}\alpha(t)}}\right)^2\right]. \tag{3.36}$$

Similar to case I, by the auto-Bäcklund transformation, the obtained solution u is the same as (3.36).

ACKNOWLEDGMENT

The authors want to express sincere thanks to Professor Gui-zhang Tu.

[1] J. W. Miles, J. Fluid Mech. **84**, 181 (1978).
[2] R. H. J. Grimshaw, Proc. R. Soc. London Ser. A **368**, 359 (1979); 377 (1979).
[3] N. Joshi, Phys. Lett. A **125**, 456 (1987).
[4] L. Abellanas and A. Galindo, Phys. Lett. A **108**, 123 (1985).
[5] Y. S. Li and B. V. Baby, "Symmetries of a KdV-type Equation," IAEG and UNESCO (1985).
[6] N. Nirmale, M. J. Vedan, and B. V. Baby, J. Math. Phys. **27**, 2640 (1986); 2644 (1986).
[7] J. Weiss, M. Tabor, and G. Carnevale, J. Math. Phys. **24**, 522 (1983); 1405 (1983).
[8] R. Conte, Phys. Lett. A **134**, 100 (1988).
[9] W. Strampp, Prog. Theor. Phys. **76**, 802 (1985).

Lie algebraic methods and solutions of linear partial differential equations

G. Dattoli, M. Richetta, G. Schettini,[a] and A. Torre
ENEA, Dip. TIB, U.S. Fisica Applicata, CRE Frascati, C.P. 65-00044-Frascati, Rome, Italy

(Received 25 July 1989; accepted for publication 5 July 1990)

In this paper, an algebraic method to obtain the solution of linear partial differential equations of the evolution type is discussed. The proposed method exploits the Lie differential operators and their matrix realization, to reduce the equation to an easily solvable generalized matrix form. Some applications to problems of specific interest are also discussed.

I. INTRODUCTION

Group theoretic methods have played an important role in the modern theory of special functions. From a historical point of view, the first deep work concerning the relationship between group representation theory and special functions is due to Cartan.[1] A more detailed and systematic use, for computational purposes, can be found in the Wigner's papers, whose elementary account is contained in his famous *Princeton Lectures* of 1955, which directly inspired the Talman book.[2] From that time, a number of milestones have been laid, the most noticeable of which is the Vilenkin work.[3] Both in Wigner's and Vilenkin's treatment, special functions are obtained as matrix elements of operators defining irreducible group representations.

The theory of special functions is strongly related to that of second-order ordinary differential equations. Lie-algebraic methods for computing eigenvalues and recurrence relation has been developed indeed (see, e.g., the papers listed in Refs. 4). In this respect, it is worth stressing that Lie introduced the groups, bearing his name, exploring the deep reasons underlying the integrability by quadratures of ordinary differential equations.

Algebraic methods dealing with partial differential equations (PDE) also have venerable roots. Schrödinger applied the method of factorization to solve the time-independent Schrödinger equation,[5] and Miller showed that this method is equivalent to the representation theory of four Lie algebras.[6] Further examples can be found in the paper by Weisner[7] and in a more systematic presentation in Miller's book.[8]

More recently Lie-algebraic methods to obtain explicit solutions to PDE have been considered by Steinberg.[9] In particular, in Ref. 9, an example generated by

$$\hat{Q} \equiv \left\{ \hat{L} = a\frac{\partial^2}{\partial x^2} + bx\frac{\partial}{\partial x} + cx^2 + \alpha\frac{\partial}{\partial x} + \beta x + \gamma; a,b,c,\alpha,\beta,\gamma \in C \right\} \tag{1.1}$$

has been discussed. Steinberg presented algebraic recipes to solve all initial value problems for operators in $\hat{Q}$. The goal of Ref. 9 was finding a solution for the following Cauchy problem:

$$\frac{\partial}{\partial t} f(x,\tau) = \hat{L} f(x,\tau), \quad f(x,0) = g(x), \tag{1.2}$$

where $g(x)$ is a sufficiently nice function and $\hat{L} \in \hat{Q}$. The solution can be written in the form

$$f(x,\tau) = e^{\hat{L}\tau} g(x). \tag{1.3}$$

Noticing that $\hat{Q}$ is a Lie algebra under the bracket operation $[\hat{L},\hat{M}] = \hat{L}\circ\hat{M} - \hat{M}\circ\hat{L} (\hat{L},\hat{M} \in \hat{Q})$, one should be able to interpret $\exp(\hat{L}\tau)$ as an element in the Lie group associated with $\hat{Q}$. The method then proceeds as follows. A suitable basis for $\hat{L}$ with generators $\hat{L}_i$ is chosen and ordering formulas of Baker, Hausdorff, Campbell, and Zassenhaus[10,11] type are used to write the evolution operator $e^{\hat{L}\tau}$ in the form

$$e^{\hat{L}\tau} = \prod_{i=1}^{n} e^{\hat{L}_i S_i(\tau)}, \tag{1.4}$$

where $S_i(\tau)$ are τ-dependent functions linked to the coefficients of (1.1).

The method proposed by Steinberg offered a unified framework to treat a wide class of Cauchy problems for linear PDE that are not of classical type (like hyperbolic or parabolic), appearing in many classical and quantum problems. In this paper, we will go a step further the method proposed in Ref. 9, presenting a technique computationally more powerful and easily generalizable to the explicitly time dependent and many variable cases.

The idea underlying this paper is very simple and can be synthetized as follows: (1) Given a PDE of the type (1.2), we identify a basis for a Lie group G; (2) since a Lie group is isomorphic to a Lie matrix group G' (Ref. 12), we utilize the matrix realization of G to write the image of our PDE in G', thus getting a straightforward solution for the characteristic ordering functions $S_i(\tau)$ appearing in (1.4); and (3) once those are obtained, we go back to the generators written in G and use simple operatorial rules to get the explicit solution of our PDE, acting as the *evolution operator* $e^{\hat{L}\tau}$ on the initial condition $g(x)$.

The method we propose has already been used in Ref. 13 to solve the Helmholtz equation in the paraxial approximation for wave propagation in selfoc media. Here, we will attempt a systematic description of the technique.

II. EXPONENTIAL OPERATORS: SOME USEFUL FORMULAS

As we have already stressed, the key point of our method is the search for an ordered product of the type (1.4),

[a] Also at ENEA, Dipartimento Fusione, CRE Frascati also at Dipartimento di Elettronica, Università di Roma, La Sapienza, Italy.

 0022-2488/90/122856-08$03.00

where $\hat{L}_i$ form the basis of a Lie algebra. The generators $\hat{L}_i$ can be written either in a matrix form or as differential operators. In this section, we will recall a few notions relevant to matrix exponent operators and to the action of exponents of differential operators on a given function $g(x)$. We will limit ourselves to very elementary properties, the interested reader is directed to Refs. 14–16 for a more complete treatment.

It happens very often in many problems of classical and quantum mechanics that the exponential $e^{\hat{a}\tau}$ must be evaluated and $\hat{a}$ is a $n\times n$ matrix, not necessarily Hermitian, with elements denoted by a_{ij}.

The Taylor expansion

$$F(\hat{a}) = e^{\hat{a}} = \sum_{n=0}^{\infty} \frac{1}{n!} (\hat{a})^n \tag{2.1}$$

is very often of limited usefulness, unless $\hat{a}$ has peculiar properties. This is indeed the case of the Pauli matrices $\hat{\sigma}_{\pm}$, $\hat{\sigma}_3$ obeying the well-known identities

$$e^{a\hat{\sigma}_+} = \begin{pmatrix} 1 & a \\ 0 & 1 \end{pmatrix}, \quad e^{b\hat{\sigma}_-} = \begin{pmatrix} 1 & 0 \\ b & 1 \end{pmatrix}, \quad e^{2c\hat{\sigma}_3} = \begin{pmatrix} e^c & 0 \\ 0 & e^{-c} \end{pmatrix}. \tag{2.2}$$

It is therefore desirable to have $F(\hat{a})$ in the closed form of a $n\times n$ matrix, rather than dealing with the not easily manageable infinite series (2.1).

To reduce $F(\hat{a})$ to a matrix form, we use the parameter differentiation technique equivalent to the Cayley–Hamilton theorem, but also useful for a time-dependent problem.

The operator function $F(\hat{a})$ acts on a n-dimensional vector space. Therefore, introducing a n-column vector ψ and a parameter τ, we can write

$$\frac{d}{d\tau} \psi(\tau) = \hat{a}\psi(\tau), \tag{2.3}$$

or, similarly,

$$\frac{d}{d\tau} \psi_i(\tau) = \sum_{j=1}^{n} a_{ij}\psi_j(\tau). \tag{2.4}$$

The above equation is of the evolution type and its solution can be cast in the form

$$\psi(\tau) = \hat{A}(\tau)\psi(0), \tag{2.5}$$

where $\hat{A}$ is a $n\times n$ matrix and $\psi(0)$ are the initial conditions of Eq. (2.3).

Substituting (2.5) in (2.3), one easily obtains

$$\frac{d}{d\tau} A_{ij} = \sum_{k=i}^{n} a_{ik}A_{kj}, \quad A_{ij}(0) = \delta_{ij}. \tag{2.6}$$

The system of linear differential equations (2.6) entirely specifies the matrix $\hat{A}(\tau)$ and, needless to say,

$$e^{\hat{a}} = \hat{A}(1). \tag{2.7}$$

For the case of a 2×2 matrix, one explicitly obtains

$$A_{11} = \left\{ \frac{1}{\sqrt{\Delta}} (a_{11} - a_{22}) \sinh\left(\frac{\sqrt{\Delta}}{2}\right) + \cosh\left(\frac{\sqrt{\Delta}}{2}\right) \right\} e^{(1/2)\mathrm{Tr}\,\hat{a}},$$

$$\frac{A_{12}}{a_{12}} = \frac{A_{21}}{a_{21}} = \frac{2}{\sqrt{\Delta}} \sinh\left(\frac{\sqrt{\Delta}}{2}\right) e^{(1/2)\mathrm{Tr}\,\hat{a}},$$

$$A_{22} = \left\{ -\frac{1}{\sqrt{\Delta}} (a_{11} - a_{22}) \sinh\left(\frac{\sqrt{\Delta}}{2}\right) + \cosh\left(\frac{\sqrt{\Delta}}{2}\right) \right\} e^{(1/2)\mathrm{Tr}\,\hat{a}}, \tag{2.8}$$

where

$$\mathrm{Tr}\,\hat{a} = a_{11} + a_{22}, \quad \Delta = (a_{11} - a_{22})^2 + 4a_{12}a_{21}. \tag{2.9}$$

It is worth stressing that, as it must be,

$$\det \hat{A} = e^{\mathrm{Tr}\,\hat{a}} \tag{2.10}$$

Since $\hat{a}$ is a nondiagonal 2×2 matrix with $\mathrm{Tr}\,\hat{a} \neq 0$ it must be treated as an element of $U(2)$.

Therefore $e^{\hat{a}}$ can also be written as the following ordered product:

$$e^{\hat{a}} = e^{2h\hat{\sigma}_3} e^{g\hat{\sigma}_+} e^{f\hat{\sigma}_-} e^{s\hat{1}}, \tag{2.11}$$

where $\hat{1}$ is the unity matrix.

Comparing (2.11) with (2.8), we obtain

$$\begin{aligned} &e^{2s} = e^{(a_{11}+a_{22})}, \\ &f = A_{21}/A_{22}, \quad g = A_{12}A_{22}\cdot e^{-(a_{11}+a_{22})} \\ &e^{-h} = A_{22}e^{-(1/2)(a_{11}+a_{22})} \end{aligned} \tag{2.12}$$

It might be worth stressing that if $\mathrm{Tr}\,\hat{a} = 0$ then $\det A = 1$ and $s = 0$, so that the basis for $\hat{A}$ becomes SU(2).

Equation (2.11) is just an ordered product, and the procedure we have outlined can be utilized as an alternative method to derive ordered forms. To give an example, we will derive within the present formalism the so-called Sack identity (see Ref. 11), namely, if $\hat{J}$ and $\hat{K}$ are operators and satisfy the following rule of commutation:

$$[\hat{J}, \hat{K}] = \lambda \hat{J}, \tag{2.13}$$

then

$$e^{\hat{J}+\hat{K}} = e^{(1/2)(1-e^{-\lambda})\hat{J}} e^{\hat{K}}. \tag{2.14}$$

An example of the algebra (2.13) is provided by the following matrix realization:

$$\hat{J} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \quad \hat{K} = \begin{pmatrix} 0 & 0 \\ 0 & \lambda \end{pmatrix}. \tag{2.15}$$

Accordingly, we get

$$e^{\hat{J}+\hat{K}} = \begin{pmatrix} 1 & (1/\lambda)(e^{\lambda}-1) \\ 0 & e^{\lambda} \end{pmatrix}. \tag{2.16}$$

To recover the identity (2.14), we write

$$e^{\hat{J}+\hat{K}} = e^{a\hat{J}} e^{\beta\hat{K}} e^{\delta\hat{1}} = \begin{pmatrix} 1 & ae^{\beta\lambda} \\ 0 & e^{\beta\lambda} \end{pmatrix}. \tag{2.17}$$

Comparing (2.17) and (2.16), the Sack identity immediately follows.

After having clarified how to handle with exponent matrix operators, it is worth, for further convenience, learning some things about the action of exponent operators in the coordinate representation on a function $g(x)$.

We consider the operator

$$\hat{L}_k = \sum_{i=1}^{m} p_{k,i}(x_1,\ldots,x_m) \frac{\partial}{\partial x_i} + q_k(x_1,\ldots,x_m), \tag{2.18}$$

where $p_{k,i}(x_1,\ldots,x_m)$ and $q_k(x_1,\ldots,x_m)$ are analytic functions. If there exist some constants $C_{rs}{}^{l}$ such that

$$[\hat{L}_r,\hat{L}_s]=\sum_{l=1}^{m} C^{l}_{r,s}\hat{L}_l, \tag{2.19}$$

we say that $\hat{L}_k$ form a basis for a Lie algebra G isomorphic to some matrix Lie algebra G' (see Ref. 12). Defining furthermore the linear combination

$$\hat{L}^{(a)}=\sum_{k=1}^{m} a_k\hat{L}_k, \tag{2.20}$$

we are interested in the action of $\exp[t\hat{L}^{(a)}]$ on an analytic function $g(x_i)$. It is almost straightforward to prove that

$$\exp[t\hat{L}^{(a)}]g(x_i)=v(t)g(X_i(t)), \tag{2.21}$$

where $X_i(t)$ and $v(t)$ are solutions of

$$\frac{d}{dt}X_i(t)=\sum_{k=1}^{m} a_k p_{k,i}(X_1(t),...,X_m(t)),$$
$$X_i(0)=x_i, \tag{2.22}$$
$$\frac{d}{dt}v(t)=v(t)\sum_{k=1}^{m} a_k q_k(X(t)), \quad v(0)=1.$$

[The proof can be easily achieved by differentiating both sides of (2.21) with respect to t and by direct substitution of $L^{(a)}$ in the final result. For a more detailed discussion, the reader is addressed to Ref. 12.] For illustrative purposes, we consider the following examples:

$$e^{a(d/dx)}g(x)=g(x+a) \tag{2.23}$$

and

$$e^{ax(d/dx)}g(x)=g(e^a x). \tag{2.24}$$

The first is trivial. The second follows from the fact that

$$\frac{d}{da}x(a)=x(a)\to x(a)=e^a x, \quad v(0)=1. \tag{2.25}$$

In the hypothesis that $e^{ax(d/dx)}$ acts on a shifted function $g(x+b)$ we get

$$e^{ax(d/dx)}g(x)=g(e^a x+b). \tag{2.26}$$

[The identity (2.26) can be easily proved, performing the change of variables $y=x+b$ and then noticing that $dy/da=y-b$.]. A further important relation is

$$e^{t(\partial^2/\partial x^2)}g(x)=\frac{1}{(4\pi t)^{1/2}}\int_{-\infty}^{+\infty}\exp\left\{-\frac{(x-y)^2}{4t}\right\}\times g(y)dy \quad (t>0). \tag{2.27}$$

The above relations are the minimal mathematical background we need to deal with algebraic methods of solution of PDE and will be largely utilized in the next sections.

III. MATRIX SOLUTIONS OF PDE: THE TIME-INDEPENDENT CASE

The method of solution we will utilize in the paper will be illustrated now with more details.

We consider the following Cauchy problem:

$$\frac{\partial}{\partial\tau}f(\tau;x_1,...,x_m)=\hat{L}(x_1,...,x_m)f(\tau;x_1,...,x_m),$$
$$f(\tau;x_1,...,x_m)\big|_{\tau=0}=g(x_1,...,x_m), \tag{3.1}$$

and

$$\hat{L}(x_1,...,x_m)=A(x_1,...,x_m)\sum_{i,j=1}^{m}\frac{\partial^2}{\partial x_i\,\partial x_j}+B(x_1,...,x_m)\times\sum_{i,j=1}^{m}x_i x_j+C(x_1,...,x_m)\sum_{i,j=1}^{m}x_i\cdot\frac{\partial}{\partial x_j}+D(x_1,...,x_m)\sum_{j=1}^{m}\frac{\partial}{\partial x_j}+E(x_1,...,x_m)\times\sum_{j=1}^{m}x_j+F(x_1,...,x_m). \tag{3.2}$$

We identify the Lie algebraic basis for $\hat{L}$ with generators $\hat{L}_i$. Using the isomorphism with a matrix algebra, we write $\hat{L}_i$ in the correspondent $m\times m$ matrix form $\hat{a}_i$ and cast Eq. (3.1) as

$$\frac{\partial}{\partial\tau}\mathbf{f}=\hat{a}\mathbf{f}, \tag{3.3}$$

where a is an $m\times m$ matrix and $\mathbf{f}$ an m-column vector. Equation (3.3) will be referred as the image equation of (3.1) in the matrix representation.

The solution of (3.3) can be written in the form

$$\mathbf{f}(\tau)=\hat{A}(\tau)\mathbf{f}(0). \tag{3.4}$$

Here, $\hat{A}(\tau)$ will be called the evolution matrix and since

$$\prod_{i=1}^{m}e^{\hat{a}_i\hat{S}_i(\tau)}=\hat{A}(\tau), \tag{3.5}$$

one can easily evaluate the ordering functions $S_i(\tau)$. Furthermore, since by virtue of the isomorphism

$$\prod_{i=1}^{m}e^{\hat{a}_i S_i(\tau)}=\prod_{i=1}^{m}e^{\hat{L}_i(x_1,...,x_m)S_i(\tau)}, \tag{3.6}$$

we can go back to the $\{\mathbf{x}\}$ realization of our Lie algebra and find the solution of (3.1) according to the standard recipe (1.3).

The method we have just illustrated can be synthesized by the following procedure. The isomorphism allows to cast Eq. (3.1) in its matrix image form (3.3), the solution is then found according to (3.4), the isomorphism allows one to go back to the original representation and find the solution of the Cauchy problem in $\{\mathbf{x}\}$ representation.

The remaining part of this section is just devoted to specific examples, the first of which is the following one-dimensional Fokker–Planck equation:

$$\frac{\partial}{\partial\tau}P(\tau;x)=\left[ax\frac{\partial}{\partial x}+\beta\frac{\partial^2}{\partial x^2}\right]P(\tau;x),$$
$$P(\tau;x)\big|_{\tau=0}=p(x). \tag{3.7}$$

It is easy to realize that the algebraic structure underlying the above equation is provided by the operators

$$\hat{k}_-=\frac{1}{2}\frac{\partial^2}{\partial x^2}, \quad \hat{k}_0=\frac{1}{2}x\frac{\partial}{\partial x}, \tag{3.8}$$

which obey the commutation relation (2.13) with $\lambda=1$. Consequently, the matrix representation for the operators $\hat{k}_-$ and $\hat{k}_0$ can be immediately written down as

$$\hat{k}_-=\begin{pmatrix}0&1\\0&0\end{pmatrix}, \quad \hat{k}_0=\begin{pmatrix}0&0\\0&1\end{pmatrix}, \tag{3.9}$$

thus specifying the matrix image of Eq. (3.7) in the form

$$\frac{\partial}{\partial\tau}P(\tau)=\begin{pmatrix}0&2a\\0&2\beta\end{pmatrix}P(\tau). \tag{3.10}$$

Accordingly, the solution to the above equation is readily obtained as

$$P(\tau) = \begin{pmatrix} 1 & B(\tau) \\ 0 & D(\tau) \end{pmatrix} P(0), \tag{3.11}$$

with

$$B = (\beta/a)(\exp(2a\tau) - 1) \quad D = \exp(2a\tau). \tag{3.12}$$

We can now go back to the x representation and get the evolution operator in the form

$$\hat{U}(\tau;\hat{x}) = e^{2h(\tau)\hat{k}_0} e^{f(\tau)\hat{k}_-} \tag{3.13}$$

with the functions h and g being

$$h(\tau) = a\tau, \quad f(\tau) = (\beta/a)(e^{2a\tau} - 1). \tag{3.14}$$

Finally using the identities (2.23)–(2.27), we easily get

$$P(\tau;x) = \hat{U}(\tau;x)p(x) = \frac{1}{[2\pi f(\tau)]^{1/2}} \int_{-\infty}^{+\infty} \exp\left\{ - \frac{[e^{h(\tau)}x - y]^2}{2f(\tau)} \right\} p(y)dy, \quad (f(\tau) > 0). \tag{3.15}$$

It is worth noticing that the operators (3.8) belong to a two-dimensional subalgebra of the SU(1,1) algebra. In order to deal with a more general case, involving the whole SU(1,1) algebra, let us consider the two-dimensional Fokker–Planck equation

$$\frac{\partial}{\partial \tau} P(\tau;x,y) = \left(axy + \beta \frac{\partial^2}{\partial x\, \partial y} \right) P(\tau;x,y), \quad P(\tau;x,y)\,|_{\tau=0} = p(x,y), \tag{3.16}$$

the relevant algebraic structure, being specified by the operators

$$\hat{k}_+ = xy, \quad \hat{k}_- = \frac{\partial^2}{\partial x\, \partial y}, \quad \hat{k}_0 = \frac{1}{2}\left(1 + x\frac{\partial}{\partial x} + y\frac{\partial}{\partial y} \right). \tag{3.17}$$

The rules of commutation

$$[\hat{k}_+, \hat{k}_-] = -2\hat{k}_0 \quad [\hat{k}_0, \hat{k}_\pm] = \pm \hat{k}_\pm \tag{3.18}$$

are immediately recognized as these relevant to the SU(1,1) algebra. Noticing that the matrix representation of the operators (3.17) is provided by

$$\hat{k}_+ = \begin{pmatrix} 0 & -1 \\ 0 & 0 \end{pmatrix}, \quad \hat{k}_- = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}, \quad \hat{k}_0 = \frac{1}{2}\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \tag{3.19}$$

we end up with the simple matrix image of Eq. (3.16)

$$\frac{\partial}{\partial \tau} P = \begin{pmatrix} 0 & -a \\ \beta & 0 \end{pmatrix} P, \tag{3.20}$$

whose solution reads

$$P(\tau) = \begin{pmatrix} \cos(\sqrt{a\beta}\,\tau) & -(a/\beta)^{1/2}\sin(\sqrt{a\beta}\,\tau) \\ (\beta/a)^{1/2}\sin(\sqrt{a\beta}\,\tau) & \cos(\sqrt{a\beta}\,\tau) \end{pmatrix} P(0). \tag{3.21}$$

Consequently, expressing the evolution operator in the ordered form,

$$\hat{U}(\tau;x,y) = e^{h(\tau)[1 + x(\partial/\partial x) + y(\partial/\partial y)]} \cdot e^{g(\tau)xy} \cdot e^{f(\tau)(\partial^2/\partial x\, \partial y)}, \tag{3.22}$$

the matrix representation (3.19) allows one to specialize the functions h, g, and f as

$$e^{-h(\tau)} = \cos(\sqrt{a\beta}\,\tau), \quad g(\tau) = -\tfrac{1}{2}(a/\beta)^{1/2}\sin(\sqrt{a\beta}\,\tau),$$
$$f(\tau) = (\beta/a)^{1/2}\tan(\sqrt{a\beta}\,\tau). \tag{3.23}$$

A further example we will discuss is rather simple and is provided by the following equation:

$$i\frac{\partial}{\partial \tau} P(\tau;x,y) = \left\{ ax\frac{\partial}{\partial y} + \beta y \frac{\partial}{\partial x} \right\} P(\tau;x,y), \quad P(\tau;x,y)\,|_{\tau=0} = p(x,y). \tag{3.24}$$

We assume that $p(x,y)$ is continuous on both x,y as well its first derivatives. An algebraic structure can be recognized using the Bergman realization of creation–annihilation operators

$$\hat{a} = \frac{\partial}{\partial x}, \quad \hat{a}^+ = x \tag{3.25}$$

and the Wigner–Schwinger representation of angular momentum

$$\hat{J}_+ = \hat{a}_1^+ \hat{a}_2, \quad \hat{J}_- = \hat{a}_1 \hat{a}_2^+, \quad \hat{J}_3 = \tfrac{1}{2}(\hat{a}_1^+ \hat{a}_1 - \hat{a}_2^+ \hat{a}_2). \tag{3.26}$$

Therefore, we can immediately identify the following SU(2)-like operators

$$\hat{J}_+ = x\frac{\partial}{\partial y}, \quad \hat{J}_- = y\frac{\partial}{\partial x}, \quad \hat{J}_3 = \frac{1}{2}\left(x\frac{\partial}{\partial x} - y\frac{\partial}{\partial y} \right). \tag{3.27}$$

[The relation of commutation are $[\hat{J}_+, \hat{J}_-] = 2\hat{J}_3, [\hat{J}_3, \hat{J}_\pm] = \pm \hat{J}_\pm$. The matrix realization of the $\hat{J}$ operators is provided by the Pauli matrices.] The solution of Eq. (3.23) can be found along the lines so far discussed thus obtaining

$$P(\tau;x,y) = \exp\left\{ -\ln[\cos(\sqrt{a\beta}\,\tau)]\left(x\frac{\partial}{\partial x} - y\frac{\partial}{\partial y} \right) \right\} \cdot \exp\left\{ -i\sqrt{\frac{a}{\beta}}\sin(2\sqrt{a\beta}\,\tau)x\frac{\partial}{\partial y} \right\} \cdot \exp\left\{ -i\sqrt{\frac{\beta}{a}}\, tg(\sqrt{a\beta}\,\tau)y\frac{\partial}{\partial x} \right\} p(x,y) = p(X(\tau), Y(\tau)), \tag{3.28}$$

where

$$\begin{pmatrix} X(\tau) \\ Y(\tau) \end{pmatrix} = \begin{pmatrix} \cos(\sqrt{a\beta}\,\tau) & -i\sqrt{\frac{a}{\beta}}\sin(\sqrt{a\beta}\,\tau) \\ -i\sqrt{\frac{\beta}{a}}\sin(\sqrt{a\beta}\,\tau) & \cos(\sqrt{a\beta}\,\tau) \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}. \tag{3.29}$$

The last example we will discuss in this section is the optical parametric amplifier equations (see Ref. 9 and references therein)

$$\frac{\partial}{\partial\tau}u(\tau;x,y) = \left\{ -\frac{1}{2}\left(\frac{\partial}{\partial x^2} - \frac{\partial^2}{\partial y^2}\right) + \left(x\frac{\partial}{\partial x} + y\frac{\partial}{\partial y}\right)\right\}u(\tau;x,y), \tag{3.30}$$

$$u(\tau;x,y)\,|_{\tau=0} = s(x,y).$$

The above equation is an almost trivial extension of Eq. (3.7) and its solution can be written as

$$u(\tau;x,y) = \frac{1}{2\pi[f_1(\tau)f_2(\tau)]^{1/2}}\int_{-\infty}^{+\infty}d\eta\int_{-\infty}^{+\infty}d\xi \times\exp\left\{-\frac{[e^{h_1(\tau)}x-\eta]^2}{2f_1(\tau)}\right\} \cdot\exp\left\{-\frac{[e^{h_2(\tau)}y-\xi]^2}{2f_2(\tau)}\right\}s(\eta,\xi), \tag{3.31}$$

with the functions $f_{1,2}$, $h_{1,2}$ specified by

$$f_1(\tau) = -\tfrac{1}{2}(e^{-2\tau}-1), \quad h_1(\tau)=\tau, \qquad f_2(\tau)=\tfrac{1}{2}(e^{2\tau}-1), \quad h_2(\tau)=\tau. \tag{3.32}$$

In this section, we have considered PDE whose underlying algebraic structure is SU(2) or SU(1,1). Within this framework, we can finally discuss the most general case of a PDE reduced to the canonical form

$$\frac{\partial}{\partial\tau}u = \{a\hat{F}_+ + \beta\hat{F}_- + \gamma\hat{F}_0 + \epsilon\hat{1}\}u, \tag{3.33}$$

where $\hat{F}$ are the generators of the simple split three-dimensional Lie algebra (SSTD). The matrix realization of $\hat{F}_{-,0}$ is the same as $\hat{k}_{-,0}$ in (3.9) while $\hat{F}_+$ is given by

$$\hat{F}_+ = \begin{pmatrix} 0 & \delta \\ 0 & 0 \end{pmatrix}. \tag{3.34}$$

According to whether $\delta = \pm 1$, the SSTD reduces to SU(2) or SU(1,1), respectively. The matrix image of (3.33) can therefore be written as

$$\frac{\partial}{\partial\tau}u = \begin{pmatrix} \epsilon+\gamma/2 & \delta a \\ \beta & \epsilon-\gamma/2 \end{pmatrix}u. \tag{3.35}$$

As a consequence, casting the evolution operator in the form

$$\hat{U}(\tau) = e^{2h(\tau)\hat{F}_0}e^{g(\tau)\hat{F}_+}e^{f(\tau)\hat{F}_-}e^{s(\tau)\hat{1}}, \tag{3.36}$$

the explicit expression of the ordering functions (h,g,f,s) can be easily derived and read [see Eqs. (2.11) and (2.12)]

$$e^{s(\tau)} = e^{\epsilon\tau},$$

$$e^{-h(\tau)} = \left\{-\frac{\gamma}{\sqrt{\Delta}}\sinh\left(\frac{\sqrt{\Delta}\tau}{2}\right)+\cosh\left(\frac{\sqrt{\Delta}\tau}{2}\right)\right\}e^{\epsilon\tau},$$

$$g(\tau) = \frac{2\delta a}{\sqrt{\Delta}}\left\{-\frac{\gamma}{\sqrt{\Delta}}\sinh^2\left(\frac{\sqrt{\Delta}\tau}{2}\right)+\frac{1}{2}\sinh\left(\sqrt{\frac{\Delta}{2}}\tau\right)\right\}e^{2\epsilon\tau},$$

$$f(\tau) = \frac{2\beta}{\sqrt{\Delta}}\left\{-\frac{\gamma}{\sqrt{\Delta}}+\operatorname{ctgh}\left(\frac{\sqrt{\Delta}\tau}{2}\right)\right\}^{-1},$$

$$\Delta = \gamma^2 + 4\delta a\beta. \tag{3.37}$$

In the next section, we will generalize the above procedure to the explicitly time-dependent case.

IV. MATRIX SOLUTIONS OF PDE: THE TIME-DEPENDENT CASE

In this section, we will consider equations of the type (3.1) in which the operator $\hat{L}$ is explicitly τ dependent. In this case, the procedure outlined in the introductory remarks of Sec. III does not change. The only difference is that the matrix elements of $\hat{A}$ are the solutions of ordinary differential equations with time-dependent coefficients. Since it is not necessary to restate the basic features of the technique, we will just discuss a few examples.

The equation we consider is a generalization of (3.7), namely,

$$\frac{\partial}{\partial\tau}P(\tau;x) = \left[a(\tau)x\frac{\partial}{\partial x}+\beta(\tau)\frac{\partial^2}{\partial x^2}+\gamma(\tau)x^2\right]P(\tau;x), \qquad P(0;x)=p(x), \tag{4.1}$$

which according to the matrix representation (3.19) of the SU(1,1) generators (3.18), can be given the matrix form

$$\frac{\partial}{\partial\tau}P(\tau) = \begin{pmatrix} a/2 & -2\gamma \\ 2\beta & -3/2a \end{pmatrix}P(\tau). \tag{4.2}$$

Introducing, as before, the transfer matrix $ABCD$ acting as

$$P(\tau) = \begin{pmatrix} A(\tau) & B(\tau) \\ C(\tau) & D(\tau) \end{pmatrix}P(0), \tag{4.3}$$

we get for the matrix elements the set of differential equations

$$\frac{\partial}{\partial\tau}A = \frac{1}{2}a(\tau)A - 2\gamma(\tau)C, \qquad \frac{\partial}{\partial\tau}C = 2\beta(\tau)A - \frac{3}{2}a(\tau)C, \qquad A(0)=1, \quad C(0)=0, \tag{4.4a}$$

$$\frac{\partial}{\partial\tau}B = \frac{1}{2}a(\tau)B - 2\gamma(\tau)D, \qquad \frac{\partial}{\partial\tau}D = 2\beta(\tau)B - \frac{3}{2}a(\tau)D, \qquad B(0)=0, \quad D(0)=1. \tag{4.4b}$$

Giving the evolution operator in the following ordered form:

$$\hat{U}(\tau;x) = \exp(s(\tau)\hat{I})\exp(2h(\tau)\hat{k}_0) \times\exp(g(\tau)\hat{k}_+)\exp(f(\tau)\hat{k}_-), \tag{4.5}$$

and comparing the matrix representation of $\hat{U}$

$$\hat{U} = e^s\begin{pmatrix} (1-fg)e^h & -ge^h \\ fe^{-h} & e^{-h} \end{pmatrix}, \tag{4.6}$$

with the $ABCD$ matrix, we can infer the link between the matrix elements of (4.3) with the ordering functions h, g, f, and s as

$$e^{2s} = AD - BC, \quad e^{-h} = D/(\sqrt{AD - BC}), \\ f = C/D, \quad g = -BD/(AD - BC). \tag{4.7}$$

Finally according to the relation (2.25)–(2.27), the solution to Eq. (4.1) specializes into

$$P(\tau;x) = e^{s+h/2} e^{1/2ge^{2h}x^2} \frac{1}{(2\pi f)^{1/2}} \\ \times \int_{-\infty}^{+\infty} e^{-(y-e^h x)^2/2f} p(y)dy \quad (f>0). \tag{4.8}$$

A further simple example is provided by Eq. (3.23) with a, β, τ-dependent functions. In this case, the solution is also easily found and has the same functional form as (3.27) with the only difference that

$$\begin{pmatrix} X(\tau) \\ Y(\tau) \end{pmatrix} = \begin{pmatrix} A(\tau) & B(\tau) \\ C(\tau) & D(\tau) \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}, \tag{4.9}$$

with A and B satisfying the following ordinary second-order differential equation:

$$\ddot{y} - (\dot{a}/a)\dot{y} + a\beta y = 0, \tag{4.10}$$

$$A(0) = 1, \quad B(0) = 0, \quad \dot{A}(0) = 0, \quad \dot{B}(0) = -ia(0),$$

and B, C

$$\ddot{y} - (\dot{\beta}/\beta)\dot{y} + a\beta y = 0, \quad C(0) = 1, \\ D(0) = 0, \quad \dot{C}(0) = -i\beta(0), \quad \dot{D}(0) = 0. \tag{4.11}$$

We believe that the examples discussed in this section are enough to stress the flexibility of the algebraic technique we propose.

For completeness sake, we conclude this section with a discussion of the more general case of a PDE reduced to the canonical form (3.33), where $(a,...,\epsilon)$ are nonsingular τ-dependent functions. In this case, the matrix image equation yields the following differential equation for the elements of the evolution matrix:

$$\begin{pmatrix} \dot{A} & \dot{B} \\ \dot{C} & \dot{D} \end{pmatrix} = \begin{pmatrix} (\gamma/2 + \epsilon)A + \delta aC & (\gamma/2 + \epsilon)B + \delta aD \\ \beta A + (\epsilon - \gamma/2)C & \beta B + (\epsilon - \gamma/2)D \end{pmatrix}. \tag{4.12}$$

On the other hand, the evolution operator (3.36) can be written in the following matrix form:

$$\hat{U} = \begin{pmatrix} (I^2 + \delta FG)\cdot H^{-1} & \delta G \\ F & H \end{pmatrix}, \tag{4.13}$$

where, for later convenience, we have defined

$$e^s = I, \quad G = ge^{h+s}, \quad F = fe^{-h+s}, \quad H = e^{-h+s}. \tag{4.14}$$

Comparing (4.12) and (4.13), we find that the functions H, G, and F obey the following first-order differential equations:

$$\dot{H} = \beta\delta G + (\epsilon - \gamma/2)H, \\ \dot{G} = (\gamma/2 + \epsilon)G + aH, \\ \dot{F} = \beta\cdot(I^2 + \delta FG)(1/H) + (\epsilon - \gamma/2)F, \tag{4.15a}$$

with initial conditions

$$H(0) = 1, \quad F(0) = G(0) = 0. \tag{4.15b}$$

The first and third equations (4.15a) can be combined to get

$$\dot{F}H - \dot{H}F = \beta I^2, \tag{4.16}$$

which amounts to saying that the determinant of $\hat{U}$ is, as it must be, $I^2 = e^{2s}$.

The evolution operator can be written in the more compact form

$$\hat{U} = \begin{pmatrix} F' & H' \\ F & H \end{pmatrix}, \tag{4.17}$$

where

$$F' = (1/\beta)[\dot{F} - (\epsilon - \gamma/2)F], \\ H' = (1/\beta)[\dot{H} - (\epsilon - \gamma/2)H]. \tag{4.18}$$

Finally, it is worth stressing that H and F satisfy the same differential equation

$$\ddot{y} + p(\tau)\dot{y} + q(\tau)y = 0, \quad p(\tau) = -\dot{\beta}/\beta - 2\epsilon, \\ q(\tau) = -\frac{\dot{\beta}}{\beta}\left(\frac{\gamma}{2} - \epsilon\right) + \frac{\dot{\gamma}}{2} - \dot{\epsilon} - \delta a\beta + \frac{\gamma^2}{4} - \epsilon^2. \tag{4.19}$$

In the next section, we will discuss examples of PDE reduced to canonical form underlying group structures with dimensionalities larger than those of SU(2) or SU(1,1).

V. CONCLUDING REMARKS

In the previous sections, we have discussed examples generated essentially by SU(2) and SU(1,1) groups. The extension to groups with higher dimensionality does not present any conceptual problem and only obvious computational difficulties may arise. This is indeed the case of the SU(3) generalization of Eq. (3.23), namely,

$$i\frac{\partial}{\partial\tau} f(\tau;x_1,x_2,x_3) = \left(\sum_{i\neq j=1}^{3} a_{i,j} x_i \frac{\partial}{\partial x_j}\right) f(\tau;x_1,x_2,x_3), \\ f(\tau;x_1,x_2,x_3)\,|_{\tau=0} = g(x_1,x_2,x_3). \tag{5.1}$$

The group structure underlying (5.1) can be recovered defining the operators[17]

$$\hat{T}_+ = x_2\frac{\partial}{\partial x_1}, \quad \hat{T}_- = x_1\frac{\partial}{\partial x_2}, \\ \hat{T}_3 = \frac{1}{2}\left(x_2\frac{\partial}{\partial x_2} - x_1\frac{\partial}{\partial x_1}\right), \\ \hat{V}_+ = x_3\frac{\partial}{\partial x_1}, \quad \hat{V}_- = x_1\frac{\partial}{\partial x_3}, \\ \hat{U}_+ = x_3\frac{\partial}{\partial x_2}, \quad \hat{U}_- = x_2\frac{\partial}{\partial x_3}, \\ \hat{Y} = \frac{1}{\sqrt{3}}\left[2x_3\frac{\partial}{\partial x_3} - x_2\frac{\partial}{\partial x_2} - x_1\frac{\partial}{\partial x_1}\right], \tag{5.2}$$

and noticing that their rules of commutation are just those of SU(3) (see Ref. 18). The solutions of Eq. (5.1) can be found along the lines discussed in the paper using the Gell–Mann and Ne'eman 3×3 matrix realization of the SU(3) generators. Skipping the rather tedious details of the derivation and referring to Ref. 17 for further comments, we just quote the final result, namely,

$$f(\tau;x_1,x_2,x_3)=g(x_1(\tau),x_2(\tau),x_3(\tau)), \tag{5.3}$$

where

$$X_i(\tau)=\sum_{j=1}^{3}A_{i,j}(\tau)x_j, \quad A_{i,j}(0)=\delta_{i,j}, \tag{5.4}$$

and any column of the matrix $\hat{A}$ satisfies the same system of first-order differential equations.

$$i\frac{d}{d\tau}\begin{pmatrix}m\\n\\p\end{pmatrix}=\begin{pmatrix}0 & a_{2,1} & a_{3,2}\\ a_{1,2} & 0 & -a_{3,1}\\ a_{2,3} & -a_{1,3} & 0\end{pmatrix}\begin{pmatrix}m\\n\\p\end{pmatrix}. \tag{5.5}$$

In analogy to what has been done in Sec. II and, for completeness sake, we will show that the technique developed in the paper can also be utilized to derive the Weyl identity (see Ref. 11). If $\hat{Q}$ and $\hat{P}$ are operators such that

$$[\hat{Q},\hat{P}]=\lambda\hat{L} \tag{5.6}$$

and $\hat{L}$ commutes with both $\hat{Q}$ and $\hat{P}$, then

$$e^{\hat{Q}+\hat{P}}=e^{\hat{Q}}e^{\hat{P}}e^{-1/2\lambda\hat{L}}. \tag{5.7}$$

The operators $\hat{Q},\hat{P},\hat{L}$ can be for example the generators of the Born–Heisenberg–Jordan group and a possible matrix realization is[2]

$$\hat{Q}=\begin{pmatrix}0&0&1\\0&0&0\\0&0&0\end{pmatrix},\quad \hat{P}=\begin{pmatrix}0&0&0\\\lambda&0&0\\0&0&0\end{pmatrix},\quad \hat{L}=\begin{pmatrix}0&0&0\\0&0&1\\0&0&0\end{pmatrix}. \tag{5.8}$$

We get, therefore,

$$e^{\hat{Q}+\hat{P}}=\begin{pmatrix}1 & 0 & 1\\ -\lambda & 1 & -\lambda/2\\ 0 & 0 & 1\end{pmatrix}. \tag{5.9}$$

The ordered form we are interested in can be cast in the form

$$e^{h\hat{Q}}e^{g\hat{P}}e^{f\hat{L}}=\begin{pmatrix}1 & 0 & h\\ -\lambda g & 1 & f\\ 0 & 0 & 1\end{pmatrix}, \tag{5.10}$$

comparing (5.10) with (5.9) the Weyl identity immediately follows.

Needless to say, the extension of the method to the case of inhomogeneous PDE is trivial. The solution of (f_0 is a known function)

$$\frac{\partial}{\partial\tau}f(\tau;x)=\hat{L}(x)f(\tau;x)+f_0(\tau;x), \quad f(\tau;x)|_{\tau=0}=g(x) \tag{5.11}$$

is obtained in the form

$$f(\tau;x)=\hat{U}(\tau;x)[g(x)+F(\tau;x)],$$
$$F(\tau;x)=\int_0^{\tau}\hat{U}^{-1}(\tau';x)f_0(\tau';x)d\tau'. \tag{5.12}$$

The algebraic technique we have presented seems to offer a powerful and feasible tool to solve a wide class of PDE. The method can be viewed as a useful alternative to more classical methods, which do not provide the same generality. Other algebraic techniques can be eventually used. For example, many operatorial identities of the Weyl, Sack–Weyl, B.C.H., and Zassenhaus type[10,11] can be directly exploited, as well as ordering theorems of the Wei–Norman (W.N.) type.[11] These methods may be useful when simple algebraic structures are involved; but become very cumbersome with increasing group dimensionality. To give an example we stress that the W.N. technique becomes particularly complicated even for the SU(3) case (see Refs. 17). The advantage of the matrix image method discussed here is that it naturally contains all the above quoted techniques, but allows us to skip most of their computational difficulties.

Needless to say, it is crucial for applying algebraic techniques in solving a PDE to identify the algebra involved. As a recipe we can suggest an evaluation of the commutators between the differential operators appearing in the equations under study, in order to find a closed structure. By so doing, it is possible to identify both the basis for the involved algebra and the relevant structure. Finally, the lowest-dimensionality matrix representation of the generators $\hat{L}_i$ can be obtained by noticing that

$$[\hat{L}_i,\hat{X}_k]=\sum_{j=1}^{m}\mathsf{L}_{i,j}\hat{X}_j, \quad k=1,...,m, \quad i=1,...,n,$$

with n and m denoting the dimensionality of the algebra and of the relevant matrix representation, respectively. The operators $\hat{X}_1,...,\hat{X}_m$ must be understood as the independent coordinates $x_1,...,x_m$ for multidimensional differential equation or as x and $\partial/\partial x$ for one-dimensional equation. For further details, the reader is addressed to Ref. 9, where a multiplication table is also provided.

A conclusive remark is in order to avoid possible misconceptions. We therefore state that the result we have obtained here, using Lie algebraic methods, could have been obtained using conventional methods. However, in practice, because of their concise and modular nature, the use of algebraic methods makes it possible to obtain results well beyond those currently available by any other method.

ACKNOWLEDGMENTS

It is pleasure to thank A. Dipace for bringing to the authors' attention the book by Srivastava and Manocha.

[1] E. Cartan, Rend. Circ. Mat. Palermo **53**, 217 (1929).
[2] J. Talman, *Special Functions* (Benjamin, New York, 1968).
[3] N. Vilenkin, *Special Functions and the Theory of Group Representation* (American Mathematical Society, Providence, RI, 1968).
[4] S. Lie, *Differentialgleichungen* (Teubner, Leipzig, 1891); L. Bianchi, *Lezioni sulla Teoria dei Gruppi Continui finiti di Trasformazioni* (Spoerri, Pisa, 1918); A. Cohen, *An Introduction to the Lie Theory of One Parameter Group* (Heath, New York, 1911); L. E. Dickson, Ann. Math. **25**, 287 (1924); P. Olver, *Applications of Lie Groups to Differential Equations* (Springer, Berlin, 1986); L. Infeld and T. Hull, Rev. Mod. Phys. **23**, 21 (1951); In this last paper algebraic methods are not explicitly mentioned, however, the factorization method used by the author can be shown to be equivalent to a group theoretic treatment.
[5] E. Schrödinger, Proc. R. Ir. Acad. Sect. A **46**, 9 (1940).
[6] W. Miller, Jr., *On Lie Algebras and Some Special Functions of Mathematical Physics* (American Mathematical Society, Providence, RI, 1964).
[7] L. Weisner, Pacific J. Math. **5**, 1033 (1955); Can. J. Math. **11**, 141 and 147 (1959).
[8] W. Miller, Jr., *Symmetry and Separation of Variables* (Addison-Wesley, London, 1977).
[9] S. Steinberg, J. Differential Eq. **26**, 404 (1977).
[10] R. M. Wilcox, J. Math. Phys. **8**, 962 (1967).

[11] G. Dattoli, A. Torre, and J. C. Gallardo, Riv. Nuovo Cimento **11**, 1 (1988).
[12] H. M. Srivastava and H. L. Manocha, *A Treatise on Generating Functions* (Ellis Horwood Lmd, Chichester, 1984).
[13] G. Dattoli, P. Di Lazzaro, and A. Torre, Nuovo Cimento B **105**, 165 (1990).
[14] J. H. Wilkinson, *The Algebraic Eigenvalue Problem* (Clarendon, Oxford, 1965).
[15] G. Dattoli, A. Torre, and M. Richetta, J. Math. Phys. **29**, 2586 (1988) (and references therein).
[16] E. G. Kalnins and W. Miller, Jr., J. Math. Phys. **15**, 1728 (1974) (and references therein).
[17] G. Dattoli, P. Di Lazzaro, and A. Torre, Phys. Rev. A **35**, 1582 (1987); G. Dattoli and A. Torre, J. Math. Phys. **31**, 236 (1990).
[18] M. Gell-Mann and Y. Ne'eman, *The Eight-fold Way* (Benjamin, New York, 1964).

A geometric approach to the path integral formalism of p-branes

Menelaos S. Kafkalidis
Department of Physics, Ohio State University, Columbus, Ohio 43210-1106

(Received 9 April 1990; accepted for publication 8 August 1990)

The configuration space for a path integral description of a p-brane is seen as a vector bundle over moduli spaces. The Einstein condition, applied to such vector bundles over compact Kähler manifolds, provides the required stability conditions. Consequently moduli spaces for such extended objects of higher dimensionality are constructed. Finally a Hermitian metric can be introduced in these moduli spaces.

I. INTRODUCTION

A geometric technique for the description of moduli spaces for the functional path integrals in string theories and other extended objects of higher dimensionality is presented. This subject seems to be of considerable interest in string theories, the statistical mechanics of multidimensional extended objects, as well as quantum gravity.

The purpose of this paper is to provide a method for the extension of the path integral formalism from string theories to extended objects of any dimension. Moreover it maintains three very advantageous points. First, the method is model independent, i.e., it can be applied to any theory that might seem appropriate. As a guiding example we consider the Polyakov string. Second, there are no limitations with respect to the dimensionality of the fundamental extended objects under consideration. This is especially important in light of the recent emergence of the p-brane and other higher-dimensional extended objects in the work of Townsend *et al.*[1-5] Third, the required *a priori* conditions for the method to be applicable are intuitively understood and are also relatively easy to verify. It must be pointed out however that the conditions are only sufficient but not necessary, that is to say we have no criterion that serves as a no-go theorem.

The questions that face us are, in principle, quite well known. Given an action S for an extended object defined classically over a world sheet (or a world volume) M, what is the configuration space $\mathcal{U}$ for the partition function Z associated with that action and its symmetry group $\mathcal{G}$? Is there a convenient description of $\mathcal{U}$? What would the corresponding path integral measure $d\mu$ be? Is there a convenient way to obtain it? The answers to these problems are, again in principle, relatively simple. The configuration space is formed as a bundle over the moduli space of M, $\mathrm{mod}(M)$. This is the analog to the process of finding the correct weights of the eigenvalues of a Hamiltonian in statistical mechanics when degeneracies are present. The measure is found from the description of the properties of one-forms in the tangent space of the moduli. This is the analog to finding dx as the integral measure in real analysis. What we strive for in this paper is to provide a physical answer to these fundamental questions, especially as they pertain to multidimensional string models.

This article is articulated into six sections. In Sec. II we present the essential steps that lead to the description of the configuration space of the path integral in terms of moduli spaces and also review the Belavin–Knizhnik theorem. In Sec. III we introduce the Einstein condition, along with some of its consequences, as the counterpart to the stability requirement necessary in the Belavin–Knizhnik theorem. We tackle the problem of the general description of the moduli space and some of its fundamental properties in Sec. IV, with the p-brane in mind. In Sec. V we propose, in passing, a way to introduce a Hermitian metric in the moduli space via an inner product of one-forms in the tangent space of the moduli. This turns out to be a convenient choice for measures in terms of applications. Finally in Sec. VI we present our concluding discussion.

II. THE TWO-DIMENSIONAL STRING

The approaches to problems of this form have traditionally elaborated on the work of Polyakov on the bosonic and fermionic strings[6] or the Green–Schwarz superstring[7,8] and have studied the space of Riemannian metrics that the base space manifold M can have, along with its associated measure.[9-12] We shall briefly review some of these results.

A. The partition function

The principal idea in the process is to start from an action of the form

$$S(\gamma,x)=\int_M \sqrt{\gamma}\gamma^{ab}\partial_a x\,\partial_b x\,d^2z, \tag{1}$$

where $\gamma^{ab}(z)$ is the metric on M and classically one would vary both $\gamma^{ab}(z)$ and $x(z)$ to obtain a Nambu–Gotto string. For such an action we attempt to construct the partition function

$$Z=\int Dx\,D\gamma\,e^{-S(\gamma,x)} \tag{2}$$

by integrating over all metrics of M denoted by $\mathcal{M}$ and all space-time embeddings denoted by $\mathcal{E}$. The problem of determining the correct configuration space $\mathcal{U}$, which is a subset of $\mathcal{M}\times\mathcal{E}$, is nontrivial because the gauge group $\mathcal{G}\equiv\mathcal{C}*\mathcal{D}$, the semi-direct product of the conformal group $\mathcal{C}$ (or the Weyl rescaling group) and the diffeomorphic group $\mathcal{D}$ (the group of coordinate transformations) leaves the action S invariant. At the same time care need be taken for the computation of the integral measure in $\mathcal{U}$.

Let us look into the first problem, the determination of

 0022-2488/90/122864-08$03.00

the configuration space. The starting point should be $\mathcal{M}\times\mathcal{E}$ from which one should eliminate the redundant degrees of freedom brought by the gauge group. As was pointed out by Alvarez[9] simply considering $\mathcal{U}=\mathcal{M}\times\mathcal{E}/\mathcal{E}*\mathcal{D}$ does not completely solve the problem of eliminating all the gauge redundancies as one might expect. In general such an excision will leave a finite number of degrees of freedom in $\mathcal{U}$, the Teichmüller parameters. As we know $\mathcal{M}/\mathcal{E}*\mathcal{D}\equiv\mathrm{mod}(M)$, the moduli space of Riemannian surfaces M. It is well known that mod (M) is a complex manifold of complex dimension

$$\dim(\mathrm{mod}(M))=\begin{cases}0, & \text{for } \nu=0,\\ 1, & \text{for } \nu=1,\\ 3\nu-3, & \text{for } \nu\geqslant 2,\end{cases} \tag{3}$$

where ν is the genus, i.e., the number of handles of the surface.

If we follow traditional conventions and divide out only the diffeomorphisms that are homotopic to the identity $\mathcal{D}_I$ then we obtain $\mathcal{M}/\mathcal{E}*\mathcal{D}_I\equiv\mathrm{t}(M)$, the Teichmüller space. Although $\mathrm{t}(M)$ is contractible, this is not necessarily the case for its quotient with the group of diffeomorphisms not homotopic to the identity, $\mathcal{D}/\mathcal{D}_I$, as this may in fact introduce nontrivial topology in $\mathcal{U}$. Therefore the general description of the configuration space $\mathcal{U}$ will be in terms of a twisted bundle over mod (M). The twists have several important consequences, especially whenever fermions are involved, since they yield discrete anomalies, but for the moment we shall not discuss such cases, interesting though they may be.

Having taken care of this problem, we can write the partition function for the string

$$Z=\int_{\mathrm{mod}(M)} d\mu\,\det'\Delta_{2\tilde{\gamma}}(\det'\Delta_{0\tilde{\gamma}})^{-13},\quad \text{for } \nu\geqslant 2, \tag{4}$$

where $d\mu$ is the Petersson–Weil measure on the moduli space, Δ_2 is a Laplacian on quadratic differentials, Δ_0 is the Laplacian on functions in M, $\tilde{\gamma}$ is a metric of constant negative curvature, and finally the prime on the determinants is a reminder that the zero modes of the Laplacians have been excluded via some regularization scheme, e.g.,

$$\det'\Delta_{0\tilde{\gamma}}=e^{-\zeta'_{\Delta}(0)}.$$

The case for genus one, the computation of Z yields the well-known result

$$Z=\frac{i}{2}\int_{M'}\frac{dz\wedge d\bar{z}}{(\mathrm{Im}\,z)^{14}|\Delta(z)|^2}, \tag{5}$$

where

$$\Delta(z)=e^{2\pi iz}\prod_{n=1}^{\infty}(1-e^{2\pi inz})^{24}$$

and the region of integration M' is specified by

$$M'=\{z:|z|\geqslant 1,\quad |\mathrm{Re}\,z|\leqslant\tfrac{1}{2},\quad \mathrm{Im}\,z>0\}.$$

As a note, observe that Eq. (5) can be expressed in terms of the theta function

$$\theta(a,z)=\sum_{n=-\infty}^{\infty}e^{2\pi ina+\pi in^2z}$$

since we have

$$\Delta(z)=e^{2\pi iz}\left(\frac{\theta(0,z)\theta(1/2,z)\theta(z/2,z)}{2}\right)^8.$$

Analogous expressions have been written for the path integral measure of the fermionic and heterotic string, where $\log(\det'\Delta)$ has been transformed into sums over appropriate geodesic curves by means of the Selberg trace formula and its supersymmetric analogue.

B. The Belavin–Knizhnik theorem

Perhaps a more useful way of looking at the problem of determining the functional measure should be a means to express $d\mu$ directly, in terms of the complex geometry of the moduli space, i.e., through properties of differential forms, rather than through spectral invariants of operators. This point of view forms the building block upon which we may extend the well-known techniques of string theories to arbitrary extended objects.

As a starting point we should review the case for the Polyakov string and then proceed to extend the results to higher dimensions. Let us begin with some results from Belavin and Knizhnik[13,14] and Mumford.[15,16] The principle idea behind this approach lies in the following nice property of the moduli space of Riemannian surfaces. As we know $\mathrm{mod}(M)$ is a complex algebraic manifold but it does not have any natural global coordinates. However the holomorphic one-forms in $\mathrm{mod}(M)$ can be described in terms of quadratic differentials on complex curves that are parametrized by the points of $\mathrm{mod}(M)$. We can find a suitable basis for the quadratic differentials $q=(q_1,q_2,\ldots,q_{3\nu-3})$ depending holomorphically on parameters from $\mathrm{mod}(M)$ and we can also have a basis for the differentials of the first kind $u=(u_1,u_2,\ldots,u_\nu)$. It has been shown in Refs. 15 and 16 that there exists a unique, up to a multiplicative constant, holomorphic function f that depends on the choices of the bases q,u such that the form

$$M_\nu=f\frac{q_1\wedge q_2\wedge\cdots\wedge q_{3\nu-3}}{(u_1\wedge u_2\wedge\cdots\wedge u_\nu)^{13}} \tag{6}$$

is a global section of an appropriate line bundle in $\mathrm{mod}(M)$. This section is meromorphic at infinity, that is to say that it is an algebraic function, has a pole of order 2, and does not have any zeros on $\mathrm{mod}(M)$. This is the celebrated Mumford form.

It was Belavin and Knizhnik in Ref. 14 that extended this statement even further by proving that the Polyakov measure on $\mathrm{mod}(M)$ is essentially the modulus squared of the Mumford form. The resulting expression for $d\mu$ is written in terms of one-forms Q_i corresponding to the q_i on $\mathrm{mod}(M)$ as

$$d\mu\propto|f|^2(-i)^\nu\frac{Q_1\wedge\bar{Q}_1\wedge\cdots\wedge Q_{3\nu-3}\wedge\bar{Q}_{3\nu-3}}{|\det\int u_i\wedge\bar{u}_j|^{13}}, \tag{7}$$

where the bar implies complex conjugation and the integral is over the Riemann surface where the u_i are defined.

Having seen this it becomes quite apparent where the advantage of such a technique lies. The computation of the path integral measure is equivalent to the description of the

Mumford form that by its definition depends *a priori* only on the *implicit global conditions*. There is another great advantage to this methodology, namely that it can be directly extended to describe the appropriate functional measure for p-dimensional strings that is presented in later sections.

There is a small technical problem to be resolved before the method presented here can be applied to extended objects in higher dimensions. The implicit conditions involved in Mumford's proof are the requirements for stability of the bundles over Riemannian surfaces. In practical terms these conditions are intractable for physicists and thus the use of this technique is relatively limited. We will, however, compensate for this technical problem by proposing instead a different but completely equivalent condition that is far more convenient to check and also physically appealing as we shall see in the immediately following section. This new condition, devoid of the handicap of stability requirements, allows us to extend the methods of Belavin and Knizhnik to the case of multidimensional strings.

III. THE EINSTEIN CONDITION

Recalling from our previous discussions that the general description of the configuration space is a *bundle* (often twisted) over $\mathrm{mod}(M)$ we should initially concentrate our efforts in studying properties of bundles. The purpose of this section is to introduce the Einstein condition, in as clear and simple way as possible, along with some of its consequences. Here lies an important contribution of this paper, for as we shall see the Einstein condition *is* the substitute for the stability requirements referred to at the end of Sec. II A. This is very important since the realization of this condition in physical terms is quite general, thus easily imposed and verified, and also it has a transparent geometric meaning which is intuitively easy to grasp.

A. Definitions and consequences

Given a holomorphic vector bundle E of rank (fiber dimension) r over a manifold M, with dimension $\dim(M)=n$, a Hermitian metric (or a Hermitian structure) h on E is a Hermitian inner product defined on the fibers of E. Bundles bestowed with such metrics are called Hermitian. On the other hand, a Hermitian metric g on the manifold M is defined as the Hermitian metric g on its tangent bundle TM.

The *mean curvature* K of a Hermitian bundle E is defined by

$$K^i_j = g^{a\bar{b}} R^{\ i}_{jab}. \tag{8}$$

Here R is the traditional curvature tensor, defined through the connection D of E. The bundle E is said to satisfy the *weak Einstein condition* with factor f provided that

$$K = f I_E \Leftrightarrow K^i_j = f\delta^i_j, \tag{9}$$

where f is a real function defined on M. This means that the mean curvature is essentially proportional to the identity. If it turns out that f is a constant, then E is said to satisfy the *Einstein condition*. A Hermitian bundle E that satisfies this condition is called an Einstein–Hermitian bundle.

The Einstein condition has far reaching consequences. Some of the simplest ones are as follows.

(i) *Every* Hermitian line bundle over a complex manifold M satisfies the weak condition, irrespective of the metric chosen for M.

(ii) A Hermitian vector bundle over a Riemannian surface satisfies the weak condition if and only if it is *projectively flat*.

(iii) The Whitney sum of two bundles E_1 and E_2 satisfies the weak condition if and only if both bundles satisfy the weak condition with the same factor f.

(iv) The tensor product of $E_1 \otimes E_2$ satisfies the weak condition with factor $f_1 + f_2$ assuming that each bundle satisfies the condition, respectively, with the appropriate factors.

(v) If the base M of the bundle E is a compact complex manifold then we have

$$\int_M c_1(E,h) \wedge \Phi^{n-1} = \frac{r}{2\pi n} \int_M f \Phi^n, \tag{10}$$

where $c_1(E,h) = -2\pi i\,\mathrm{tr}(R)$ is the first Chern form of E with respect to the metric and Φ is the *fundamental two-form*, also known as the *Kähler two-form*, which can be expressed in terms of local coordinates on M by $\Phi = \sqrt{-1} g_{i\bar{j}} dz^i \wedge d\bar{z}^j$.

However, the important consequence of the Einstein condition, of interest to physicists, is the fact that *every* irreducible Einstein–Hermitian bundle over a compact Kähler manifold is stable.[17-19] As already mentioned the concept of stability of a bundle is not new. It is exactly this idea, pioneered by Mumford, that led to the Belavin–Knizhnik theorem when applied to bundles over Riemann surfaces.

The need for stability arises from the fact that the moduli space of holomorphic structures need not be connected in general (i.e., it's non-Hausdorff) unless we are limited to stable (or semi-stable) structures (see Ref. 20 for example). Stability then is a *de facto* necessity for the existence of "good," i.e., physically interesting moduli spaces. If the moduli spaces are not connected then the entire concept of perturbation theory fails since there will certainly exist areas of the moduli, i.e., acceptable metrics, which will be impossible to reach by infinitesimal deformations of "fiducial metrics." We shall not pursue here the details of the algebraic theory but we shall present the fundamental concept of stability.

The definition of Φ-stability (or Φ-semistability) is as follows. Assume that $\mathcal{S}$ is a coherent sheaf over a compact Kähler manifold (M,g) of dimension n and Φ is the Kähler form of (M,g). Since the manifold in Kähler Φ is a closed $(1,1)$-form on M. The first Chern class of $\mathcal{S}$ is defined by the first Chern class of the determinant bundle of $\mathcal{S}$, i.e.,

$$c_1(\mathcal{S}) = c_1(\det \mathcal{S}). \tag{11}$$

This is also represented by a real closed $(1,1)$-form on M. The Φ-degree of $\mathcal{S}$ is defined as [see also Eq. (10)]

$$\deg(\mathcal{S}) = \int_M c_1(\mathcal{S}) \wedge \Phi^{n-1}. \tag{12}$$

The degree to rank ratio of $\mathcal{S}$ is defined, in direct analogy to the degree to rank ratio for bundles, as

$$\mu(\mathcal{S}) = \deg(\mathcal{S})/\operatorname{rank}(\mathcal{S}). \tag{13}$$

The sheaf $\mathcal{S}$ is said to be Φ-semistable provided that for every coherent subsheaf $\mathcal{S}'$ of positive rank we have

$$\mu(\mathcal{S}') \leqslant \mu(\mathcal{S}).$$

If, moreover, it happens that $\mu(\mathcal{S}') < \mu(\mathcal{S})$ for all subsheaves with $0 < \operatorname{rank}(\mathcal{S}') < \operatorname{rank}(\mathcal{S})$ then $\mathcal{S}$ is said to be Φ-stable. By analogy, a vector bundle E over M is said to be Φ-stable or Φ-semistable if the sheaf $\mathcal{S}(E)$ of germs of holomorphic sections of E is respectively Φ-stable or Φ-semistable.

B. Criteria for stability

The mathematical definition of stability is of very little use to physicists as it would be extremely cumbersome to check in every case. This is the reason why we have introduced the Einstein condition in its place. We present a convenient criterion, stated in the form of a theorem, that guarantees the stability of bundles under two conditions that are quite general and relatively easy to verify. The stability theorem provides the necessary requirements for stability of bundles as well as a useful decomposition property. It can be put in the following way.

Theorem 1: If E is an Einstein–Hermitian bundle, with a Hermitian structure h, satisfying the Einstein condition with factor f, over a compact Kähler manifold M, with metric g, then E is Φ-semistable and it is a direct sum

$$(E,h) = \bigoplus_k (E_k,h_k)$$

of Φ-stable Einstein–Hermitian vector bundles (E_k,h_k) with the same factor f as (E,h).

A crucial feature here is that there is *no* limitation as to the dimensionality of the base manifold M. We shall use this to our advantage in the following sections.

This result is analogous to several algebraic theorems proven by Takemoto,[21] Bogomolov,[22] and Gieseker,[23] which generalize Mumford's results to the case of the stability of sheaves, over algebraic surfaces of higher dimensions.

IV. THE MODULI SPACE

Stability of Einstein–Hermitian bundles is not limited to two-dimensional base manifolds and here lies the importance of the Einstein condition. Everything that up to now has been applied to the Polyakov string can now be applied to the case of p-dimensional extended objects. The extension of the action in Eq. (1) to the p-brane case is trivial and the partition function is still given by Eq. (2), the differences being that the world sheet M is now a world volume and that the gauge group $\mathcal{G}$ may not be the same as for the two-dimensional string.

Since the discussion that follows does *not* depend on a particular model for a p-brane we shall explicitly refrain from choosing one over any other. It must be pointed out, once again, that this technique applies to any string model that we may come up with, two-dimensional or p-dimensional, and consequently we do not wish to pose any limitations in that respect. We shall however play favorites and proceed with p-branes in mind, since there is a need to properly describe the steps necessary for the construction of the partition function.

The essential fact remains that the configuration space for the partition function is still going to be described as a bundle over mod (M) and consequently it should be treated as such. Also we need to keep in mind that we will require a description for the tangent space of the moduli, since there lies the solution to the problem of finding the appropriate integration measure. These questions are answered in the following sections.

As already mentioned, the first nontrivial problem namely the connectedness of the moduli space for the p-branes case has been eliminated. The reason of course is that the connectedness of the moduli space is guaranteed as long as stability is maintained, that is as long as the Einstein condition is satisfied. We shall therefore assume from the outset that we can always impose the Einstein condition. Consequently we can proceed to the next step, namely the description of the moduli space, pertinent to the general p-dimensional string, and its properties.

It has been shown[24,25] that the moduli space of Einstein–Hermitian connections is open in the moduli space of holomorphic structures. Moreover it has a natural complex structure, very much analogous to the Dolbeault complex of holomorphic structures, which can be utilized to describe some of its important properties in terms of cohomology groups of the bundle.

A. Definitions and basic notions

Consider E a C^∞ complex vector bundle of rank r over a complex manifold M and $\mathrm{GL}(E)$ the group of automorphisms of E, that is the group of transformations that induce the identity transformation on the base M, and $\mathcal{D}_0(E)$ the set of linear maps between differential one-forms $D_0 : A^0(E) \to A^{0,1}(E)$ that satisfy

$$D_0(ab) = d_0 a \cdot b + a \cdot D_0 b, \quad \text{for } a \in A^0(E) \text{ and } b \in A^0.$$

All of these maps can be generalized to higher-order (p,q)-forms $D_0 : A^{p,q}(E) \to A^{p,q+1}(E)$, $p,q \geqslant 0$ such that:

$$D_0(\alpha,\beta) = d_0\alpha \wedge \beta + (-1)^{r+s}\alpha \wedge D_0\beta,$$
$$\text{for } \alpha \in A^{r,s}, \beta \in A^{p,q}(E). \tag{14}$$

Consider the subset $\mathcal{H}_0(E) \subseteq \mathcal{D}_0(E)$ containing connections that satisfy the condition $D_0 \circ D_0 = 0$. This set can be thought of as the set of holomorphic structures in E. The group $\mathrm{GL}(E)$ acts on $\mathcal{D}_0(E)$ in the usual way

$$D_0 \to D'_0 = \tau^{-1} \circ D_0 \circ \tau = D_0 + \tau^{-1} \circ d_0 \tau,$$
$$D_0 \in \mathcal{D}_0(E) \text{ and } \tau \in \mathrm{GL}(E). \tag{15}$$

Equivalence of holomorphic structures implies that they belong in the same $\mathrm{GL}(E)$ orbit. Then one has $\mathcal{H}_0(E)/\mathrm{GL}(E)$ as the moduli space of holomorphic structures in E.

Consider now a Hermitian structure h in E and take $U(E) \subset \mathrm{GL}(E)$ to be the unitary automorphisms of E. The connections $D(E) \in \mathcal{D}(E)$ that are induced by the structure will be of the form $D : A^0(E) \to A^1(E)$ and their action is

$$D(ab) = da\cdot b + a\cdot Db, \text{ for } a\in A^0(E) \text{ and } b\in A^0,$$
$$d(h(a_1,a_2)) = h(Da_1,a_2) + h(a_1,Da_2),$$
$$\text{for } a_1,a_2\in A^0(E). \tag{16}$$

All these connections can be extended uniquely to the case of general (p,q)-forms in the sense that $D:A^{p,q}(E) \to A^{p+1,q}(E) + A^{p,q+1}(E)$ which obey

$$D(\alpha\wedge\beta) = d\alpha\wedge\beta + (-1)^{r+s}\alpha\wedge D\beta,$$
$$\text{for } \beta\in A^{p,q}(E),\ \alpha\in A^{r,s}. \tag{17}$$

One can also naturally decompose the action of the connection among (0,1)- and (1,0)-forms by having $D = D_0 + D_1$, where $D_0:A^0(E)\to A^{0,1}(E)$ and $D_1:A^0(E)\to A^{1,0}(E)$ and thus obtain the maps

$$\mathscr{D}(E)\to\mathscr{D}_0(E), \quad \text{when } D\to D_0. \tag{18}$$

As a matter of fact, if we know $D_0\in\mathscr{D}_0(E)$ then we can easily determine the remainder D_1 from the relation

$$d_0(h(a_1,a_2)) = h(D_0a_1,a_2) + h(a_1,D_1a_2),$$
$$\text{for } a_1,a_2\in A^0(E) \tag{19}$$

and by its construction $D = D_0 + D_1\in\mathscr{D}(E)$. We can always consider $\mathscr{D}(E)$ as an infinite-dimensional affine space isomorphic to $A^1(\mathrm{End}(E))$, i.e., the space of one-forms with values in the bundle $\mathrm{End}(E)$. Again in a completely analogous fashion as previously we can define the subspace of connections $\mathscr{H}(E)\subset\mathscr{D}(E):D = D_0 + D_1$ such that their (0,1) components satisfy $D_0\circ D_0 = 0$. Equivalently we can say

$$\mathscr{H}_0(E) = \{D_0\in\mathscr{D}_0(E) \text{ with } D_0\circ D_0 = 0\}.$$
$$\mathscr{H}(E) = \{D\in\mathscr{D}(E) \text{ with } D_0\in\mathscr{H}_0(E)\}.$$

As one might expect $\mathscr{H}(E)$ and $\mathscr{H}_0(E)$ and $\mathscr{D}(E)$ and $\mathscr{D}_0(E)$ are closely linked. We can see this in the following way. Suppose $D\in\mathscr{H}(E)$. As we know D_0 defines a unique structure h in E such that $D_0 = d_0$ and consequently this connection is the Hermitian connection of E with respect to the structure. On the other hand, each holomorphic structure $D_0\in\mathscr{H}_0(E)$ yields a unique connection $D = D_0 + D_1\in\mathscr{H}(E)$ which of course is the Hermitian connection of E with respect to D_0. Essentially there is a map that links the two spaces $b:\mathscr{H}(E)\leftrightarrow\mathscr{H}_0(E)$.

The action of $U(E)$ on $\mathscr{D}(E)$ is in analogy with Eq. (15)

$$D\to D' = u^{-1}\circ D\circ u = D + u^{-1}\circ du,$$
$$\text{for } D\in\mathscr{D}(E) \text{ and } u\in U(E) \tag{20}$$

and will leave $\mathscr{H}(E)$ invariant. However note that since there is a bijection between $\mathscr{D}_0(E)$ and $\mathscr{D}(E)$ the action of $\mathrm{GL}(E)$ on $\mathscr{D}_0(E)$ must have some bearing on $\mathscr{D}(E)$. Such a correspondence is established via

$$D\to D' = \tau^*\circ D_1\circ(\tau^*)^{-1} + \tau^{-1}\circ D_0\circ\tau,$$
$$\text{for } D\in\mathscr{D}(E) \text{ and } \tau\in\mathrm{GL}(E), \tag{21}$$

where τ^* is the adjoint of τ in the sense that $h(a_1,\tau^*a_2) = h(\tau a_1,a_2)$. We then have the following proposition.

Proposition 2: The unitary subgroup $\mathrm{U}(E)$ contains automorphisms of E that are parallel with respect to D and consequently it is naturally isomorphic to the centralizer of the unitary group $\mathrm{U}(r)$ of the holonomy group of the connection D. Consequently it is compact.

Notice that if the holonomy group of D is irreducible then $\mathrm{U}(E)$ is essentially the group of scalar multiplication by $\theta\in C:|\theta| = 1$ and thus it is isomorphic to $\mathrm{U}(1) = \theta I_E;\ \theta\in C:|\theta| = 1$. This is much smaller than the center of $\mathrm{U}(E)$ in general, which would contain automorphisms of the form ϕI_E, where ϕ are functions with $|\phi| = 1$.

We also propose that the action of $\mathrm{U}(E)$ on $\mathscr{D}(E)$ or $\mathscr{H}(E)$ is proper. Consequently the quotient spaces $\mathscr{D}(E)/\mathrm{U}(E)$ and $\mathscr{H}(E)/\mathrm{U}(E)$ are certainly going to be Hausdorff.

With the added assumption that the base manifold M is a compact Kähler one, with Kähler metric g, we can consider $\mathscr{E}(E)\subset\mathscr{H}(E)$ which will contain Einstein–Hermitian connections, that is to say

$$\mathscr{E}(E) = \{D\in\mathscr{H}(E):K(D) = \zeta I_E\}, \tag{22}$$

where ζ is a constant and $K(D)$ is the mean curvature of the connection. In general, if $D\in\mathscr{D}(E)$ and $u\in\mathrm{U}(E)$ then

$$K(D') = u^{-1}\circ K(D)\circ u \tag{23}$$

and in addition

$$R(D') = (u^{-1}\circ D\circ u)\circ(u^{-1}\circ D\circ u)$$
$$= u^{-1}\circ R(D)\circ u \tag{24}$$

and consequently one can see that the action of $\mathrm{U}(E)$ on $\mathscr{D}(E)$ will leave $\mathscr{E}(E)$ invariant. By defining the quotient space $\mathscr{E}(E)/\mathrm{U}(E)$ as the moduli space of Einstein–Hermitian structures in E we have the following proposition.

Proposition 3: The moduli space $\mathscr{E}(E)/\mathrm{U}(E)$ of Einstein–Hermitian structures in E is Hausdorff and can be injected into the moduli space $\mathscr{H}_0(E)/\mathrm{GL}(E)$ of holomorphic structures in E.

B. Properties of moduli

Structures, i.e., inner products, defined on bundles are not always the most convenient concepts to handle. However structures in bundles induce connections, a concept that we are much more familiar with, and as a result there is always a strong link between the properties of structures and connections. It should come as no surprise then that the previous theorem can be extended to read as follows.

Theorem 4: The moduli space $\mathscr{E}(E)/\mathrm{U}(E)$ of Einstein–Hermitian connections in E is open in the moduli space $\mathscr{H}_0(E)/\mathrm{GL}(E)$ of holomorphic structures in E.

This theorem allows us to work with Einstein–Hermitian bundles and connections, as opposed to structures, knowing *a priori* that the moduli space will be well defined.

In addition to this theorem, it has been shown[25] that the moduli space of Einstein–Hermitian connections is in effect describable in terms of cohomology groups of the bundle. This is the analogous statement to the Belavin–Knizhnik theorem for the Polyakov string.

Theorem 5: The moduli space $\mathscr{E}_i(E)/\mathrm{U}(E)$ of irreducible Einstein–Hermitian connections in E is a complex analytic space, nonsingular at the point D if

$H^2(M,\mathrm{End}^0(E_{D_0}) = 0$ and its tangent space at such a point is isomorphic to the sheaf cohomology group $H^1(M,\mathrm{End}(E_{D_0}))$.

Please note that $\mathrm{End}^0(E_{D_0}) \subset \mathrm{End}(E_{D_0})$ is the space of traceless endomorphisms of the bundle. Also note that there is no mention of the dimensionality of the base manifold.

The physical implications here are obvious. The previous theorem describes the moduli space, essentially via the connection, from the symmetry group of the action. In addition the description of the tangent space of the moduli in terms of cohomology groups of the bundle is a welcome bonus, since it will help us determine the path integral measure associated with the partition function.

It is important to know the dimensionality of the moduli spaces. At least for the cases where the base manifold M happens to be compact and Kähler, with a Kähler form Φ with a Hermitian vector bundle E of rank r over it, then the moduli space $\mathscr{E}(E)/\mathrm{U}(E)$ is a nonsingular Kähler manifold if $\int c_1(M) \wedge \Phi \geqslant 0$, where $c_1(M)$ is the first Chern class of M. The dimension of the moduli space, assuming that it is not empty, can be easily computed from the Riemann–Roch formula and is equal to one of the following:

(i) $2rc_2(E) - (r-1)c_1(E)^2 + r^2(h^{0,1} - 1) + 1$ if $\int_M c_1(M) \wedge \Phi \geqslant 0$ and the canonical line bundle L_M is nontrivial,

(ii) $2rc_2(E) - (r-1)c_1(E)^2 + r^2(h^{0,1} - 2) + 2$ if L_M is trivial,

(iii) $2rc_2(E) - (r-1)c_1(E)^2 - r^2 + 1$ if $c_1(M) > 0$,

where $c_p(E)$, $p = 1,2$ are the Chern classes of E and $h^{p,q}$ are the appropriate Hodge numbers. As we know there is a simple relation between these, the Betti numbers b_s and the Euler characteristic χ of a manifold, namely,

$$\chi = \sum(-1)^s b_s = \sum(1)^{p+q} h^{p,q}.$$

Note that if L_M is indeed trivial then M is either a torus with $h^{0,1} = 2$ or a K3 surface with $h^{0,1} = 0$ and then the moduli space can have a holomorphic symplectic structure. These are in fact the only classes of surfaces with symplectic structures in two dimensions. In higher dimensions though, should one require the imposition of a symplectic structure, several different possibilities must be investigated. The reason symplectic structures are mentioned here is due to the fact that they appear to be quite useful in compactification schemes at least as far as string models are concerned.

V. THE METRIC

At this point we wish to define a Hermitian metric in the moduli space. This is not an essential feature of the method we present and thus the reader may directly proceed to the next section. However, it is a convenient choice to be made in the case of applications. The essential point of this section lies in the fact that the tangent space of the moduli is isomorphic to $H^1(\mathrm{End}(E_{D_0}))$ and consequently an introduction of a Hermitian inner product in the latter will generate a Hermitian metric in the former.

A. Definitions

Let us begin from the L^2_k space of one-forms over M with values in the skew Hermitian endomorphisms of E. Call

$$\Lambda = L^2_k(A^1(\mathrm{End}(E,h))), \quad \text{for } k > n. \tag{25}$$

For any $\psi \in \Lambda$ we can always write the usual decomposition in (0,1) and (1,0)-forms as

$$\psi = \psi_0 + \psi_1, \tag{26}$$

where ψ_0 is a (0,1)-form and ψ_1 is a (1,0)-form. Then the condition that ψ is skew-Hermitian can be written as

$$\bar{\psi}^t = -\psi \Leftrightarrow \psi_1 = -\bar{\psi}_0^t. \tag{27}$$

We can easily define a complex structure J on Λ as

$$J\psi = i\psi_0 - i\psi_1, \tag{28}$$

so that we have naturally decomposed Λ in two parts

$$K = L^2_k(A^{0,1}(\mathrm{End}(E))), \quad \bar{K} = L^2_k(A^{1,0}(\mathrm{End}(E))). \tag{29}$$

Obviously the natural isomorphism from $\Lambda \to K$ (or $\bar{K}$) is obtained by setting

$$\psi \to \psi_0 \text{ (or } \psi_1). \tag{30}$$

In order to define a Hermitian inner product in K consider the local case of the inner product $(u,v): u,v \in K$. Expressing this in terms of an orthonormal basis $(e_1, e_2, \ldots, e_r)$ of the fiber of E and an orthonormal frame $(m_1, m_2, \ldots, m_n)$ of the base M we can write

$$u(e_i) = \sum_{j,k} u^j_{ik} \bar{m}^k e_j = u^j_{ik} \bar{m}^k e_j,$$

$$v(e_i) = \sum_{j,k} v^j_{ik} \bar{m}^k e_j = v^j_{ik} \bar{m}^k e_j, \tag{31}$$

and thus the inner product can be written as

$$\langle u,v \rangle = \sum_{ijk} u^j_{ik} \bar{v}^j_{ik}. \tag{32}$$

A more useful expression for the inner product of one-forms, in terms of their traces, can be given. Recalling that $v^\dagger = \bar{v}^t$ and that

$$\mathrm{tr}(u \wedge v^\dagger) = -\sum_{ijkl} u^j_{ik} \bar{v}^j_{il} m^l \wedge \bar{m}^k, \tag{33}$$

we can write the inner product as

$$\langle u,v \rangle = Q^\dagger(-i\,\mathrm{tr}(u \wedge v^\dagger)), \tag{34}$$

where $Q^\dagger$ is the adjoint operator of $Q = (\Phi \wedge \cdot)$, which is nothing more than the exterior multiplication with the Kähler form. On the other hand, we also have

$$-in\,\mathrm{tr}(u \wedge v^\dagger) \wedge \Phi^{n-1} = \langle u,v \rangle \Phi^n, \tag{35}$$

and consequently we can define a Hermitian inner product h' on K by

$$h'(u,v) = \int_M \langle u,v \rangle \Phi^n$$

$$= -i \int_M n\,\mathrm{tr}(u \wedge v^\dagger) \wedge \Phi^{n-1}. \tag{36}$$

From this we can see the corresponding inner product h in Λ to be

$$h(u,v) = h'(u_0,v_0) + h'(v_0,u_0)$$
$$= -i\int_M n\,\mathrm{tr}(u_0 \wedge v_0^\dagger + v_0 \wedge u_0^\dagger) \wedge \Phi^{n-1}$$
$$= -i\int_M n\,\mathrm{tr}(u_0 \wedge v_1 - u_1 \wedge v_0) \wedge \Phi^{n-1}, \quad (37)$$

and the two-form ω corresponding to this inner product would be

$$\omega(u,v) = \int_M n\,\mathrm{tr}(u \wedge v) \wedge \Phi^{n-1}. \quad (38)$$

It so happens[26] that this inner product will in fact induce a Kähler metric on $\mathscr{E}(E)/\mathrm{U}(E)$, by applying it to harmonic forms, $u,v \in H^1(M,\mathrm{End}(E_{D_0}))$. Recall that by Theorem 5 H^1 is isomorphic to the tangent space of the moduli space.

Also noteworthy is the fact that the real part of the Kähler metric, namely the Riemannian metric on the moduli, will depend on the Riemannian metric on M but it will *not* depend on its complex structure. For example suppose that M is a compact Riemannian manifold with a holonomy group such that M is Ricci flat and Kähler with respect to any complex structure consistent with its holonomy. The Riemannian metric on the moduli is independent of the complex structure we choose on M despite the fact that the complex structure and the Kähler structure of the moduli do in fact depend on the complex structure on M.

The Riemannian metric is easily obtained by applying Eq. (37) to harmonic forms u,v in the real tangent space $\mathscr{E}(E)/\mathrm{U}(E)$. Then

$$g(u,v) = h(u_0,v_0) + \overline{h(u_0,v_0)}$$
$$= \int_M (\langle u_0,v_0\rangle + \overline{\langle u_0,v_0\rangle})\Phi^n$$
$$= \int_M \langle u,v\rangle \Phi^n, \quad (39)$$

where $\langle u,v\rangle$ is defined by the Riemannian metric and Φ^n would be the volume element of M.

VI. CONCLUDING REMARKS

Let us recapitulate the new and important contributions of this paper for the path integral formulation of p-brane theory. First, as we have seen, the configuration space can be described as a vector bundle over $\mathrm{mod}(M)$. It becomes apparent then that the introduction of the Einstein condition is the *only* requirement needed and this greatly facilitates our study of $\mathscr{Q}$, the configuration space, over which the partition function is defined. Second, the path integral measure can be found in the tangent bundle of $\mathrm{mod}(M)$. This has been described in terms of the cohomology of the bundle. Third, we have at our disposal the convenience to introduce a Hermitian metric in the moduli space.

In this method it is important to recognize the fact that the only *a priori* condition we required depends on the implicit *global* topology of the string (or membrane) and not on spectral invariants of local operators. The condition is both intuitive and simple as well as convenient to check for any particular string model we wish to consider. It is also interesting to note that this technique applies equally well to 2-D or p-dimensional string theories.

As an example, consider the cases of supersymmetric 2-D string theories, where the base manifold of the string was proposed to be Calabi-Yau,[27,28] i.e., a compact Kähler manifold with SU(3) holonomy. The Einstein condition is trivially satisfied for *all* these models. In addition we know already what the measure looks like by the Belavin–Knizhnik construction. On the other hand, there are other physical considerations that cause serious problems. These have nothing to do with the applicability of the Einstein condition but stem from purely dynamical considerations pertaining to the model. We know now that considering such base spaces, on which to write a supersymmetric σ-model action, will lead to a nonvanishing beta function in fourth- and fifth-order perturbation calculations[29,30] the physical implication here being that the theory does not have asymptotically free behavior from its inception. There is no known result of similar nature that blemishes the extension of this model to the case of a supersymmetric p-dimensional membrane.

Also noteworthy is the fact that this description of the moduli space for the partition function, as well as the tangent space, for multidimensional extended strings comes to complement the classification of p-dimensional extended supersymmetric objects by Townsend *et al.*,[1,2] where the Eilenberg–Chevalley cohomology of the super-translation group Σ was considered. The pertinent bosonic, or supersymmetric, action would lead to a definition for an appropriate connection and consequently provide the metric on the base M. Despite the fact that no attempts were made therein to write a path integral formalism for such models, it is quite obvious that the Einstein condition must be verified *a priori* before we can proceed with the computation of the partition function or more importantly consider any perturbative expansions for multiloop calculations. Once again it must be mentioned that there is no reason to disqualify a certain string model on the grounds that it does not satisfy the Einstein condition. However, our analysis in such cases will become quite complicated and we should be extremely careful in our construction of moduli spaces appropriate for the partition functions and their corresponding measures.

This is of course only the beginning of the investigation and there is a need for further development and application of these techniques to potentially interesting multidimensional string models that will be presented elsewhere.[31]

ACKNOWLEDGMENTS

The author wishes to thank his Ph.D. thesis advisor Dr. Walter W. Wada for his continual support and many stimulating discussions.

[1] A. Achúcarro, J. Evans, P. K. Townsend, and D. L. Wiltshire, Phys. Lett. B **198**, 441–446 (1987).
[2] J. A. de Azcárraga and P. K. Townsend, Phys. Rev. Lett. **62**, 2579–2582 (1989).
[3] E. Bergshoeff, E. Sezgin, and P. K. Townsend, Phys. Lett B **189**, 75–78 (1987).
[4] J. Hughes, J. Liu, and J. Polchinski, Phys. Lett B **180**, 370–374 (1986).
[5] P. K. Townsend, Phys. Lett. B **202**, 53–57 (1988).
[6] A. Polyakov, Phys. Lett. B **103**, 207–210 (1981); **103**, 211-213 (1981).

[7] M. B. Green, J. H. Schwarz, and E. Witten, *Superstring Theory* (Cambridge U.P., Cambridge, 1987).
[8] J. Schwarz, Phys. Rep. **89**, 223 (1982).
[9] O. Alvarez, Nucl. Phys. B **216**, 125–184 (1983).
[10] L. Alvarez-Gaumé and P. Nelson, CERN Preprint CERN-TH 4615/86.
[11] E. D'Hoker and D. H. Phong, Nucl. Phys. B **269**, 205–234 (1986); **278**, 225–241 (1986); **292**, 317–329 (1987); Princeton preprint PUPT-1039 (1988).
[12] G. Moore and P. Nelson, Nucl. Phys. B **266**, 58–74 (1986).
[13] A. A. Beilinson and Y. I. Manin, Commun. Math. Phys. **107**, 359–376 (1986).
[14] A. A. Belavin and V. G. Knizhnic, Sov. Phys. JETP **64**, 214–228 (1986).
[15] D. Mumford, Enseign. Math. **23**, 39–100 (1977).
[16] D. Mumford and J. Fogarty, *Geometric Invariant Theory* (Springer Verlag, Berlin, 1982), 2nd ed.
[17] S. Kobayashi, Proc. Jpn. Math. Acad. **58**, 158–162 (1982); Proc. Jpn. Acad. **62**, 21–24 (1986); Nagoya Math. J. **101**, 37–54 (1986).
[18] M. Lübke, Manuscripta Math. **42**, 247–252 (1983).
[19] C. Okonek, M. Schneider, and H. Spindler, *Vector Bundles on Complex Projective Spaces*, Progress in Math. Vol. 3 (Birkhäuser, Boston, 1980).
[20] V. A. Norton, Indiana Univ. Math. J. **28**, 365-387 (1979).
[21] F. Takemoto, Nagoya Math. J. **47**, 29–48 (1973); **52**, 173–195 (1973).
[22] F. A. Bogomolov, Math. USSR Izv. **13**, 499–555 (1978); Proc. Int. Congr. Math. Helsinki, 517–524 (1978).
[23] D. Gieseker, Ann. Math. **106**, 45-60 (1977); Am. J. Math. **101**, 77–85 (1979).
[24] H. J. Kim, "Curves and Holomorphic Vector Bundles," Ph.D thesis, Univ. of California, Berkeley, 1985 (unpublished).
[25] H. J. Kim, Math. Z. **195**, 143–150 (1987).
[26] Y-T. Siu, *Lectures on Hermite-Einstein Metrics for Stable Bundles and Kahler-Einstein Metrics*, DMV Seminar (Birkhäuser, Basel, 1987), Vol. 8.
[27] P. Candelas, G. Horowitz, E. Strominger, and E. Witten, Nucl. Phys. B **258**, 46–74 (1985).
[28] E. Witten, E. Nucl. Phys. B **258**, 75–100 (1985).
[29] M. T. Grisaru, A. E. M. van de Ven, and D. Zanon, Phys. Lett. B **173**, 423–428 (1986); Nucl. Phys. B **277**, 388–408 (1986); **277**, 409–428 (1986).
[30] M. T. Grisaru, D. I. Kazakov, and D. Zanon, Nucl. Phys. B **287**, 189–204 (1987).
[31] M. Kafkalidis, in preparation.

Bäcklund transformations for surfaces in Minkowski space

Bennett Palmer
Technische Universitat Berlin, Fachbereich 3—Mathematik, Strasse des 17 Juni, 135, 1 Berlin 12, West Germany

(Received 13 January 1989; accepted for publication 27 June 1990)

A Bäcklund transformation is constructed between spacelike surfaces of constant negative curvature and timelike surfaces of constant negative curvature in three-dimensional Minkowksi space. The transformation gives a differential geometric interpretation to a Bäcklund transformation between the elliptic sine-Gordon equation and the elliptic sinh-Gordon equation studied by Leibbrandt [J. Math. Phys. **19** (1978)].

I. INTRODUCTION

The classical Bäcklund transformation[1] for surfaces of constant negative curvature in the Euclidean space $\mathbb{E}^3$ has been an object of study for over 100 years. Under this transformation, an infinite family of constant curvature surfaces can be produced from a given one. In recent times, interest in this transformation has persisted due to its connection with the sine-Gordon equation and its associated soliton theory.

In this paper, we investigate an analogous transformation for surfaces in the Minkowski space $\mathbb{E}_1^3$.

Theorem I: Let $M \subset \mathbb{E}_1^3$ be a spacelike surface with constant negative curvature. Corresponding to M is an infinite one-parameter family of timelike surfaces in $\mathbb{E}_1^3$ with constant negative curvature.

One difference between our approach and the classical result is that elliptic equations are utilized in place of hyperbolic ones. For the spacelike surface, the Gauss equation is the elliptic sinh-Gordon equation (EShGE)

$$\Delta u = \sinh u \tag{1.1}$$

and for timelike surfaces the elliptic sine-Gordon equation (ESGE)

$$\Delta v = \sin v, \tag{1.2}$$

where $\Delta = \partial_{xx}^2 + \partial_{yy}^2$. A transformation between solutions of (1.1) and (1.2) has been studied previously by Leibbrandt[2] in connection with the theory of superconductors. Theorem I can be considered a geometrization of this result.

Throughout this paper, all surfaces are assumed to be of class C^∞. We use the abbreviations sl for spacelike and tl for timelike. The symbols ch, sh, etc. are used to denote the hyperbolic cosine, sine, etc.

II. PRELIMINARIES

Throughout this section, we use the summation convention. Let

$$X: M \to \mathbb{E}_1^3$$

be a sl immersion of a surface M. On a neighborhood in M select a frame $\{e_j\}_{j=1,2,3}$ with $\{e_j\}_{j=1,2}$ orthonormal and e_3 a normal to M with $\langle e_3, e_3 \rangle = -1$. Define the dual coframe by $w^i(e_j) = \delta_j^i$. We have

$$dX = e_j w^j. \tag{2.1}$$

Differentiation of e_j defines an so (2.2) valued one-form $w = (w_i^j)$ by

$$de_i = e_j w_i^j. \tag{2.2}$$

By considering the equation $d\,dX = 0$, one obtains

$$dw^i = w^j \wedge w_j^i, \quad i = 1, 2, 3. \tag{2.3}$$

In particular, when $i = 3$, we get the "symmetry equation"

$$w^1 \wedge w_1^3 + w^2 \wedge w_2^3 = 0, \tag{2.4}$$

along M. From the equations $d\,de_j = 0, j = 1, 2, 3$, we obtain the equation of Gauss

$$dw_1^2 = w_1^3 \wedge w_3^2 \tag{2.5}$$

and the Codazzi equations

$$dw_j^3 = w_j^i \wedge w_i^3, \quad j = 1, 2. \tag{2.6}$$

Cartan's Lemma and (2.4) allows one to set

$$w_3^j = -h_{ij} w^i, \tag{2.7}$$

with $h_{ij} = h_{ji}$. The (Gauss) curvature K of M is given by

$$dw_1^2 = -K w^1 \wedge w^2 = w_1^3 \wedge w_3^2, \tag{2.8}$$

so that $-K = \det(h_{ij})$. The mean curvature h of M is defined by

$$2h w^1 \wedge w^2 = w_3^1 \wedge w^2 + w^1 \wedge w_3^2, \tag{2.9}$$

so that $2h = -(h_{11} + h_{22})$. Note the sign convention.

In what follows it will be useful to introduce isothermal coordinates on M. Denote such a coordinate by $z = x + iy$, the induced metric on M is expressed

$$ds^2 = e^\rho |dz|^2. \tag{2.10}$$

Differentiating the normal field $\eta \equiv e_3$ on M defines Φ by

$$\eta_z = -\Phi e^{-\rho} X_{\bar{z}} - h X_z \tag{2.11}$$

where the subscripts z, $\bar{z}$ are, respectively, the complex derivatives $\frac{1}{2}(\partial_x \mp i\partial_y)$. The structural equations of the immersion X are

$$\begin{aligned} X_{zz} &= \rho_z X_z - (\Phi/2)\eta, \\ X_{z\bar{z}} &= -h(e^\rho/2)\eta, \\ \eta_z &= -hX_z - \Phi e^{-\rho} X_{\bar{z}}. \end{aligned} \tag{2.12}$$

By taking the third derivatives with respect to z and $\bar{z}$ of the first two equations and equating the mixed partials, one obtains the integrability equations of Gauss

$$|\Phi|^2 e^{-2\rho} = h^2 - 2e^{-\rho}\rho_{z\bar{z}} \tag{2.13}$$

and Codazzi

 0022-2488/90/122872-04$03.00

$$\Phi_{\bar{z}} = e^{\rho} h_z. \tag{2.14}$$

The quantity $Q \equiv \Phi\, dz^2$ defines an invariant differential on M. Under change of isothermal coordinate $z = z(\zeta)$, Q transforms according to

$$Q = \Phi\, dz^2 = \Phi\left(\frac{dz}{d\zeta}\right)^2 ds^2. \tag{2.15}$$

Note that by (2.14), Q is a holomorphic quadratic differential when $h = \text{const}$.

The map $d\eta$ defines a self-adjoint linear transformation of the tangent space at each point in M. Denote its eigenvalues by $-k_i$, $i = 1, 2$. It follows from the last equation of (2.12) that

$$e^{-\rho}|\Phi| = (k_1 - k_2)/2 \tag{2.16}$$

and

$$h = -[(k_1 + k_2)/2]. \tag{2.17}$$

We remark that given a coordinate z as above one may obtain an orthonormal frame by setting

$$e_1 - ie_2 = e^{-\rho/2} X_z. \tag{2.18}$$

For timelike immersions,

$$f: T \to \mathbb{E}_1^3,$$

we introduce an adapted frame on a neighborhood in T such that e_3 is the unit normal and the induced metric is

$$\tau^1 \otimes \tau^1 - \tau^2 \otimes \tau^2, \tag{2.19}$$

where τ^i is dual to e_j. We have the equations

$$de_j = \tau_j^k e_k, \quad j = 1, 2, 3, \tag{2.20}$$

with $\tau_1^2 = \tau_2^1$. The Gauss curvature K is then given by

$$d\tau_1^2 = -K\tau^1 \wedge \tau^2. \tag{2.21}$$

III. PROOF OF THEOREM I

Let

$$Y: M \to \mathbb{E}_1^3 \tag{3.1}$$

be a spacelike immersion. Let λ be a constant and consider a second immersion

$$\tilde{Y} = Y + \lambda(\cos\theta e_i + \sin\theta e_2). \tag{3.2}$$

We seek a function θ on M such that

(i) the displacement

$$\epsilon_1 \equiv \cos\theta e_1 + \sin\theta e_2$$

is tangent to $\tilde{M} \equiv \tilde{Y}(M)$ and

(ii) the normal $\tilde{\epsilon}_3$ to $\tilde{M}$ is of the form

$$\tilde{\epsilon}_3 = \operatorname{ch}\phi(-\sin\theta e_1 + \cos\theta e_2) + \operatorname{sh}\phi e_3, \tag{3.3}$$

for some constant ϕ.

The condition (ii) prescribes the angle between M and $\tilde{M}$ to be a constant. We compute

$$\begin{aligned} d\tilde{Y} = {} & [w^1 - \lambda \sin\theta(d\theta + w_1^2)]e_1 \\ & + [w^2 + \lambda\cos\theta(d\theta + w_1^2)]e_2 \\ & + [\lambda\cos\theta w_1^3 + \lambda\sin\theta w_2^3]e_3. \end{aligned} \tag{3.4}$$

The condition $0 = \langle d\tilde{Y}, \tilde{\epsilon}_3 \rangle$ yields

$$\begin{aligned} -\lambda\, d\theta = {} & \cos\theta(w^2 - \lambda\tanh\phi w_1^3) \\ & - \sin\theta(w^1 + \lambda\tanh\phi w_2^3) + \lambda w_1^2. \end{aligned} \tag{3.5}$$

Using the structural equations on M the integrability equation $d\, d\theta = 0$ yields, after much computation,

$$w^1 \wedge w^2 + \lambda^2 \tanh^2\phi w_1^3 \wedge w_2^3 + \lambda^2 K w^1 \wedge w^2 = 0, \tag{3.6}$$

which implies

$$\operatorname{ch}^2\phi + \lambda^2 K = 0. \tag{3.7}$$

Proposition: There exists a surface satisfying (i) and (ii) iff. $K \equiv \text{const} < 0$ on M.

We now investigate the surface $\tilde{M}$. Assume for convenience $k \equiv -1$. Clearly the condition (ii) implies $\tilde{M}$ is timelike. Let θ be the solution obtained above and change coframe on M according to

$$\sigma^1 + i\sigma^2 = e^{-i\theta}(w^1 + iw^2). \tag{3.8}$$

With the obvious notation, it is easily checked that

$$\sigma_1^2 = w_1^2 + d\theta \tag{3.9}$$

and

$$\sigma_1^3 + i\sigma_2^3 = e^{-i\theta}(w_1^3 + iw_2^3). \tag{3.10}$$

Using (3.8)–(3.10) and taking for convenience $\lambda = +\operatorname{ch}\phi$, (3.5) becomes

$$\sigma^2 = \operatorname{ch}\phi\sigma_1^2 + \operatorname{sh}\phi\sigma_1^3.$$

Let ϵ_j be the frame on M dual to σ^j. Then, $\tilde{Y} = Y + \operatorname{ch}\phi\epsilon_1$ and computation yields

$$d\tilde{Y} = \tilde{\epsilon}_1\tilde{\epsilon}^1 + \tilde{\epsilon}_2\tilde{\sigma}^2,$$

where

$$\begin{aligned} & \tilde{\epsilon}_1 = \epsilon_1, \\ & \tilde{\epsilon}_2 = \operatorname{sh}\phi\epsilon_2 + \operatorname{ch}\phi\eta, \quad \epsilon_3 = \operatorname{ch}\phi\epsilon_2 + \operatorname{sh}\phi\eta, \\ & \tilde{\sigma}^1 = \sigma^1, \quad \tilde{\sigma}^2 = \sigma_1^3, \quad \tilde{\sigma}_1^2 = -\operatorname{sh}\phi\sigma_1^2 + \operatorname{ch}\phi\sigma_1^3, \\ & \tilde{\sigma}_1^3 = \operatorname{ch}\phi\sigma_1^2 - \operatorname{sh}\phi\sigma_1^3, \quad \tilde{\sigma}_2^3 = \sigma_2^3. \end{aligned} \tag{3.11}$$

It is straightforward to check that

$$d\tilde{\sigma}_1^2 = \tilde{\sigma}^1 \wedge \tilde{\sigma}^2,$$

which implies the curvature $\tilde{K}$ of $\tilde{M} \equiv \tilde{Y}(M)$ satisfies $\tilde{K} \equiv -1$ and Theorem I follows.

IV. COORDINATE-DEPENDENT CONSEQUENCES

In this section, we investigate coordinate-dependent consequences of Sec. III. In order to make the choice of coordinate as natural as possible, we parametrize M as a parallel to a sl constant mean curvature surface.

Proposition: Let

$$X: \Sigma \to \mathbb{E}_1^3 \tag{4.1}$$

be an umbilic, free sl immersion with constant mean curvature $h \neq 0$. Define the parallel surface

$$Y = X + (1/2h)\eta. \tag{4.2}$$

Then $M \equiv Y(\Sigma)$ is a sl surface with constant curvature $\tilde{K} \equiv -4h^2$.

Proof: Let $\{E_j\}$ be frame on a neighborhood in Σ with

dual coframe $\{\alpha^i\}$ and connection forms $\{\alpha^i_1\}$. Define $\{w^i\}_{j=1,2}$ by

$$dY = e_1 w^1 + e_2 w^2. \tag{4.3}$$

It follows from (4.2) that

$$w^j = \alpha^j + (2h)^{-1}\alpha^j_3, \quad j = 1, 2. \tag{4.4}$$

It is also easily verified using the Gauss and Codazzi equations on Σ that (with the obvious notation)

$$w^j_1 = \alpha^j_1, \quad 1 \leqslant i, j \leqslant 3. \tag{4.5}$$

Therefore,

$$d\alpha^2_1 = -K\alpha^1 \wedge \alpha^2 = -\tilde{K} w^1 \wedge w^2. \tag{4.6}$$

However,

$$\begin{aligned} w^1 \wedge w^2 &= (\alpha^1 + (2h)^{-1}\alpha^1_3) \wedge (\alpha^2 + (2h)^{-1}\alpha^2_3) \\ &= -(2h)^{-2} K \alpha^1 \wedge \alpha^2 \end{aligned}$$

and the result follows.

Let Σ be as in the proposition. We assume for convenience that $h \equiv -\frac{1}{2}$. The general case can be obtained by rescaling X. Locally it is possible an isothermal coordinate $\zeta = x + iy$ so that

$$Q = \tfrac{1}{2}\, d\zeta^2.$$

To see this replace an arbitrary isothermal z by

$$\zeta(z) \equiv \int_{z_0}^{z} (2\Phi(w))^{1/2}\, dw,$$

for some branch of the square root and some fixed $z_0 \in \Sigma$. The transformation rule for Q yields

$$Q = \Phi\, dz^2 = \Phi\left(\frac{dz}{d\zeta}\right)^2 d\zeta^2 = \frac{1}{2}\, d\zeta^2.$$

Write the metric induced on Σ as

$$ds^2 = e^{\rho} |d\zeta|^2.$$

The choice of coordinate ζ yields a frame

$$E_1 = e^{-\rho/2} X_x, \quad E_2 = e^{-\rho/2} X_y, \tag{4.7}$$

which is tangent to the principle directions, so that

$$\alpha^j_3 = k_j \alpha^j, \quad j = 1, 2. \tag{4.8}$$

By (1.16)

$$|\Phi| e^{-\rho} = (h^2 + K)^{1/2} = (k_1 - k_2)/2. \tag{4.9}$$

This implies, since $\Phi \equiv \frac{1}{2}$, that

$$(k_1 - k_2)/2 = \tfrac{1}{2} e^{-\rho} \tag{4.10}$$

and since

$$\tfrac{1}{2} = -h = (k_1 + k_2)/2 \tag{4.11}$$

we have

$$k_1 = \tfrac{1}{2}(1 + e^{-\rho}), \quad k_2 = \tfrac{1}{2}(1 - e^{-\rho}). \tag{4.12}$$

Therefore by (4.8)

$$w^1 = \alpha^1 - \alpha^1_3 = \operatorname{sh}(\rho/2) dx, \tag{4.13}$$

$$w^2 = \alpha^2 - \alpha^2_3 = \operatorname{ch}(\rho/2) dy, \tag{4.14}$$

and so the metric $d\sigma^2$ induced on M is

$$d\sigma^2 = \operatorname{sh}^2(\rho/2) dx^2 + \operatorname{ch}^2(\rho/2) dy^2.$$

A standard formula for the Gauss curvature yields

$$\tilde{K} = -1 \Leftrightarrow \rho_{xx} + \rho_{yy} = \operatorname{sh}\rho \quad \text{(EShGE)}.$$

Using (4.12), one finds that

$$w^1_3 = \operatorname{ch}(\rho/2) dx, \quad w^2_3 = \operatorname{sh}(\rho/2) dy.$$

From (3.11) and (4.2), it follows that

$$\tilde{\sigma}^1 = \sigma^1 = \cos\theta \operatorname{sh}(\rho/2) dx + \sin\theta \operatorname{ch}(\rho/2) dy,$$

$$\tilde{\sigma}^2 = \sigma^3_1 = \cos\theta \operatorname{ch}(\rho/2) dx + \sin\theta \operatorname{sh}(\rho/2) dy.$$

Finally, changing the coframe $(\tilde{\sigma}^1, \tilde{\sigma}^2)$ by a Lorentz transformation,

$$\begin{pmatrix} \tau^1 \\ \tau^2 \end{pmatrix} = \begin{pmatrix} \operatorname{ch}\dfrac{\rho}{2} & -\operatorname{sh}\dfrac{\rho}{2} \\ -\operatorname{sh}\dfrac{\rho}{2} & \operatorname{ch}\dfrac{\rho}{2} \end{pmatrix} \begin{pmatrix} \tilde{\sigma}^1 \\ \tilde{\sigma}^2 \end{pmatrix}$$

brings the pseudometric $d\tilde{\sigma}^2$ induced on $\tilde{M}$ into the form

$$\begin{aligned} d\tilde{\sigma}^2 &= \tau^1 \otimes \tau^1 - \tau^2 \otimes \tau^2 \\ &= \sin^2\theta\, dy^2 + \cos^2\theta\, dx^2. \end{aligned}$$

As a consequence

$$\tilde{K} = -1 \Leftrightarrow \theta_{xx} + \theta_{yy} = \tfrac{1}{2}\sin 2\theta \quad \text{(ESGE)}.$$

Theorem II: The Bäcklund transformation of Theorem I induces a Bäcklund transformation between solutions of (EShGE) and (ESGE).

A transformation appearing in the work of Leibbrandt[2] connects the two equations mentioned above. We will now show how to recover Liebbrandt's transformation from Theorems I and II.

Let Σ be as above with $h \equiv -\frac{1}{2}$ and ζ chosen so that $\Phi \equiv \frac{1}{2}$. We have

$$\eta_\zeta = \tfrac{1}{2} X_\zeta - \tfrac{1}{2} e^{-\rho} X_{\bar{\zeta}}.$$

The immersion $\tilde{Y}$ in (3.2) can be expressed

$$Y = X - \eta - \operatorname{ch}\phi (e^{i\theta - \rho/2} X_\zeta + e^{-i\theta - \rho/2} X_{\bar{\zeta}}),$$

while the normal $\tilde{\epsilon}_3$ is given by

$$\tilde{\epsilon} = \operatorname{ch}\phi (i e^{i\theta - \rho/2} X_\zeta + i e^{-i\theta - \rho/2} X_{\bar{\zeta}}) + \operatorname{sh}\phi \eta. \tag{4.15}$$

A tedious computation using the structural equations, (2.12) shows that the equation (3.5) is equivalent to

$$\operatorname{ch}\phi \left(\theta_\zeta - i\frac{\rho\zeta}{2}\right) = \left(\frac{1}{2} - \frac{i}{2}\operatorname{sh}\phi\right) \sin\left(\theta + i\frac{\rho}{2}\right),$$

setting $\cos\lambda = \operatorname{sch}\phi$, $\sin\lambda = \tanh\phi$, this may be written

$$\left(\theta - i\frac{\rho}{2}\right)\zeta = \frac{1}{2} e^{-i\lambda} \sin\left(\theta + \frac{i\rho}{2}\right). \tag{4.16}$$

This is essentially the transformation appearing in Ref. 1.

Remark: Leibbrandt's "procedure of harmonic functions" arises naturally by varying the choice of isothermal coordinate on Σ.

Remark: On Σ, θ is an invariantly defined function satisfying

$$\Delta_\Sigma \theta = (k_1 - k_2)\sin 2\theta,$$

where Δ_Σ is the Laplace–Beltrami operator of Σ.

Remark: It can be shown that $\tilde{\epsilon}_3$ defined by (4.15) with θ satisfying (4.16) is a harmonic map of Σ into $S_1^2 \equiv \{w \in \mathbb{E}_1^3 \,|\, \langle w,w\rangle = +1\}$. Since constant mean curvature, sl surfaces can be recovered from harmonic maps into $H^2 = \{w \in \mathbb{E}_1^3 \,|\, \langle w,w\rangle = -1\}$, there is an induced Bäcklund transformation between the corresponding harmonic maps.

ACKNOWLEDGMENT

The author would like to thank G. Leibbrandt for his encouragement.

[1] L. P. Eisenhardt, *A Treatise on the Differential Geometry of Curves and Surfaces* (Dover, New York, 1909).

[2] G. Leibbrandt, "Soliton solutions of the elliptic sine-cosine equations by means of harmonic functions," J. Math. Phys. **19**, (1978).

Extremal properties of Synge's world function and discrete geometry

Yu. A. Rylov
Institute for Problems in Mechanics, Academy of Sciences of the USSR, Prospect Vernadskogo 101, Moscow 117526, USSR

(Received 18 August 1989; accepted for publication 28 February 1990)

Properties of σ space [a set Ω of points P with a real function $\sigma(P,P')$ given on Ω] are investigated. A continuity of the set Ω is not necessary and, generally, geometry is discrete. The properties of the world function σ are investigated. At certain (extremal) world function properties the σ space is shown to be a subset of points of Euclidean space or Riemannian space. The presented approach has the peculiarity that no operation other than the function σ is given on σ space. In particular, all such operations as linear operation over vectors, constructing lines and planes, and dimension of the space are expressed through the world function σ and only through it (if it is extremal). A violation of the σ-space extremality leads to going out beyond the frames of Riemannian geometry (lines are substituted by tubes of lines, etc.). The presented approach can be useful in quantum gravitation, string models, and other problems, where the properties of the event space at small distances are important.

I. INTRODUCTION

Event space (pseudo-Euclidean or Riemannian) is used to consider a set of two independent structures: (i) linear vector space (or manifold in the case of a Riemannian space) and (ii) metric space. Linear vector space properties are used for constructing straight lines, planes, and for determination of space dimensionality. Metric space properties are used for determining distances, volumes, etc. The concept of continuity is important for describing linear vector space.

In general, the hierarchy of concepts can be represented as follows: (i) the manifold, which includes continuity and affine properties, including space dimensionality; (ii) metric properties; and (iii) the topological type of the space (i.e., whether it is topologically equivalent to a plane, cylinder, sphere, etc).

Another approach to the description of event space properties is possible. This approach uses only the structure of metric space. Linear space and its properties are considered not as a new additional structure, but as a structure generated by the metric space.

Let us illustrate the structure of metric space in the example of three-dimensional proper Euclidean space E_3, where the distance $d(P_1,P_2)$ between the points P_1 and P_2 with the Cartesian coordinates $\mathbf{x}_1$ and $\mathbf{x}_2$ is defined in the conventional way:

$$d(P_1,P_2)=d(\mathbf{x}_1,\mathbf{x}_2)=\sqrt{(\mathbf{x}_1-\mathbf{x}_2)^2}=|\mathbf{x}_1-\mathbf{x}_2|. \tag{1.1}$$

It is easy to see that

$$d(P_1,P)+d(P,P_2)=d(P_1,P_2), \tag{1.2}$$

considered as an equation for the point P at fixed P_1 and P_2 ($P_1\neq P_2$), determines an intercept $\mathscr{L}_{[P_1P_2]}$ of the straight line $\mathscr{L}_{P_1P_2}$ between the points P_1, P_2. Equation (1.2), considered as an equation for P_2 and fixed $P_1\neq P$, determines a ray of the straight line, passing through the points P_1, $P\in E_3$. Hence, solving the algebraic equation of type (1.2), one can construct a straight line passing through two points.

If one can construct straight lines, then one can construct two-dimensional planes, etc. and determine the space dimensionality. Practically, all properties (except continuity) of the Euclidean space, which are usually described in terms of the linear space, can be obtained from properties of the distance function d of the metric space.

Such an approach is possible only in the case when the distance $d(P_1,P_2)$ has specific (extremal) properties. For instance, if expression (1.1) for the distance $d(\mathbf{x}_1,\mathbf{x}_2)$ is substituted by

$$d(\mathbf{x}_1,\mathbf{x}_2)=\sqrt{(\mathbf{x}_1-\mathbf{x}_2)^2-a|\mathbf{x}_1-\mathbf{x}_2|/(1+a|\mathbf{x}_1-\mathbf{x}_2|^2)}, \tag{1.3}$$

where a is a nonvanishing constant, then the set of points P satisfying Eq. (1.2) forms a two-dimensional cigar-shaped surface. In the case when $0<a\ll 1$ the transversal size of the "cigar" of unit length is much less than its longitudinal size. Then the cigar distinguishes itself slightly from the intercept of the straight line.

In the conventional approach the geodesic (straight) line is defined as a curve of an extremal (in the given case of the shortest) length. It is natural that the definition of the curve must be given and the space dimensionality concept must be defined.

In our approach the geodesic (straight) line is merely a set of points which is determined by the form of the distance function $d(\mathbf{x}_1,\mathbf{x}_2)$. The circumstance that the line is one-dimensional (but not two-dimensional, for instance) is conditioned by specific properties of the function d. There is no necessity to introduce concepts such as continuity, curve, and dimensionality. The distance function can be given on *any set of points.*

Practically, it is more convenient to use another function

$$\sigma(P_1,P_2)=\tfrac{1}{2}d^2(P_1,P_2) \tag{1.4}$$

instead of the function d. The function σ will be referred to as the Synge world function or merely the σ function. The properties of the world function that provide, in general, the

 0022-2488/90/122876-15$03.00

degeneration of the many-dimensional surface (1.2) into the one-dimensional geodesic will be referred to as extremal properties. The present paper is devoted to investigating extremal properties of the σ function.

The real event space description in terms of the σ function realizes an intuitive conception that the event space is described by and only by an interval between any pair of events. This circumstance, together with the permission of discrete space, can conveniently occur at the event space quantization. Finally, the potential violation of the σ-function extremality at small distances leads to the fact that a pointlike particle is described by a world tube (but not by a one-dimensional world line): It is associated with the string model in elementary particle theory.

Apparently, the world function was first introduced for the Riemannian space description by Ruse[1,2] and Synge.[3] In gravitation theory the world function is denoted by different symbols: by Ω in Synge;[4] by G in Rylov;[5,6] where the world function is used for the two-metric formalism, by I in Gorelik,[7] which is devoted to the introduction of special coordinate systems on the base of the world function. In current quantum gravitation[8,9] the symbol σ is used. This designation will be used further.

In the presentation of event space properties in terms of the world function the hierarchy of suppositions is as follows: (i) determination of the world function, (ii) world function extremality, and (iii) continuity properties.

The world function extremality is equivalent to the Euclidean axiom: One and only one straight line passes through two different points, one and only one plane passes through three different points that do not lie on one straight line, etc. If the world function is extremal, then the space dimensionality and topological type of the space (plane, cylinder, sphere, etc.) can be determined in some cases independently of the continuity property, for instance, in the case of a discrete space.

The circumstance that the continuity is nonessential can be useful in quantum gravitation because in a certain sense the quantization is a substitution of continuous variables by discrete ones.

Section II is devoted to the introduction of σ space and its extremality. In Section III the Euclidean space properties are described in terms of the σ function. In Sec. IV we investigate to what extent the σ space extremality determines its properties. In Sec. V curtailed tubes are introduced and the manifold is described in terms of the σ function. Section VI is devoted to the description of Riemannian space in terms of the σ function. In Sec. VII some violations of the σ-function extremality are investigated.

II. σ SPACE

Definition 2.1: The σ space V is a set Ω of points P with a real function σ of two points P and Q given on Ω, where the function σ has the properties

$$\sigma(P,Q)=\sigma(Q,P), \quad \sigma(P,P)=0, \quad P,Q\in\Omega. \tag{2.1}$$

The function σ will be referred to as the world function, or merely as the σ function.

The interval S between the points P and Q is defined as

$$S(P,Q)=\sqrt{2\sigma(P,Q}$$
$$=\begin{cases}|\sqrt{2\sigma(P,Q)}|, & \sigma(P,Q)\geqslant 0,\\ i|\sqrt{2\sigma(P,Q)}|, & \sigma(P,Q)<0,\end{cases}$$
$$P,Q\in\Omega \tag{2.2}$$

Let us introduce more a real function of the three points P_0, P_1, P_2:

$$\Gamma(P_0,P_1,P_2)\equiv\sigma(P_0,P_1)+\sigma(P_0,P_2)$$
$$-\sigma(P_1,P_2), \quad P_0,P_1,P_2\in\Omega. \tag{2.3}$$

The functions $\sigma(P,Q)$, $S(P,Q)$ are symmetric with respect to the arguments P and Q: $\Gamma(P_0,P_1,P_2)$ is symmetric only with respect to the arguments P_1 and P_2.

It is easy to see that any subset $\Omega'\subset\Omega$ of the σ-space points is a σ space.

Let us introduce the designations

$$^k\mathscr{P}^n=\{P_k,P_{k+1},\ldots,P_n\}, \quad {}^k\mathscr{P}^n_e={}^k\mathscr{P}^n\setminus\{P_e\},$$
$$\mathscr{P}^n\equiv{}^0\mathscr{P}^n, \quad \mathscr{P}^n_e=\mathscr{P}^n\setminus\{P_e\}, \tag{2.4}$$

and define a real function F_n of $n+1$ points $\mathscr{P}^n\subset\Omega$:

$$F_n(\mathscr{P}^n)=\det\|\Gamma(P_0,P_i,P_k)\|, \quad i,k=1,2,\ldots,n. \tag{2.5}$$

One can show that as a result of Eq. (2.3) $F_n(\mathscr{P}^n)$ is symmetric with respect to any pair of points P_i *and* P_k $(i,k=0,1,\ldots,n)$.

The meaning of the functions σ, Γ, F_n can be understood most easily in the example of D-dimensional proper Euclidean space, considering it as a σ space with the σ function

$$\sigma(P,Q)=\sigma(x,y)=\frac{1}{2}\sum_{i,k=1}^{D}g_{ik}(x^i-y^i)(x^k-y^k), \tag{2.6}$$

where $\{x^i\}$ and $\{y^i\}$ are contravariant coordinates of the points P and Q, respectively, in some coordinate system K. Here $g_{ik}=\text{const}$, $i,k=1,2,\ldots,D$ is the metric tensor in the coordinate system K, which is formed by $D+1$ points P_0, $P_1,\ldots,P_D$ with the point P_O as an origin of K. The vectors

$$\mathbf{e}_i=\mathbf{P_0P}_i,$$
$$|\mathbf{e}_i|=|\mathbf{PP}_i|=\sqrt{2\sigma(P_0,P_i)}, \quad i=1,2,\ldots,D \tag{2.7}$$

are directed along coordinate axes of the coordinate system K. Then using the cosine theorem and Eqs. (2.7) and (2.3), it is easy to verify that

$$g_{ik}=(\mathbf{e}_i,\mathbf{e}_k)$$
$$=\tfrac{1}{2}[|\mathbf{P_0P}_i|^2+|P_0P_k|^2-|P_iP_k|^2]$$
$$=\Gamma(P_0,P_i,P_k), \quad i,k=0,1,\ldots,D. \tag{2.8}$$

The function (2.5),

$$F_n(\mathscr{P}^n)=\det\|g_{ik}\|\ .$$
$$=\det\|(\mathbf{P_0P}_i,\mathbf{P_0P}_k)\|=(n!V_n(\mathscr{P}^n))^2, \tag{2.9}$$

is the Gram determinant and $V_n(\mathscr{P}^n)$ is the volume of $(n+1)$ edron with vortices at the points $\mathscr{P}^n$. In proper Euclidean space the condition

$$F_n(\mathscr{P}^n)\neq 0 \tag{2.10}$$

is the necessary and sufficient condition of the linear independence of n vectors $\mathbf{e}_i = P_0P_i$, $i = 0,1,...n$ and that of the fact that $n+1$ points $\mathscr{P}^n$ do not lie on one $(n-1)$-dimensional plane.

Definition 2.2: The $n+1$ point basis $\mathscr{P}^n$ is $n+1$ points $P_i \in \Omega$ $(i = 0,1,...,n)$ that satisfy condition (2.10).

The point basis $\mathscr{P}^n$ in the Euclidean space is associated with the basis of n vectors $\mathbf{e}_i = \mathbf{P}_0\mathbf{P}_i$ $(i = 1,2,...,n)$ in an n-dimensional plane $\mathscr{L}(\mathscr{P}^n)$ containing the points $\mathscr{P}^n$.

Definition 2.3: The line tube (or merely the tube) $\mathscr{T}_n$ of the nth order formed by the basis $\mathscr{P}^n$ is a set of points $P \in \Omega$:

$$\mathscr{T}_{\mathscr{P}^n} = \mathscr{T}(\mathscr{P}^n) = \{P \mid F_{n+1}(P,\mathscr{P}^n) = 0\}, \quad F_n(\mathscr{P}^n) \neq 0. \tag{2.11}$$

Definition 2.4: Section $\mathscr{S}_{n;P}$ of the tube $\mathscr{T}(\mathscr{P}^n)$ at the point $P \in \mathscr{T}(\mathscr{P}^n)$ is the set of points $P' \in \Omega$:

$$\mathscr{S}_{n;P}(\mathscr{T}(\mathscr{P}^n)) = \left\{P' \middle| F_{n+1}(P'\mathscr{P}^n) = 0 \overset{l=n}{\underset{l=0}{\wedge}} \sigma(P_l,P') = \sigma(P_l,P)\right\}, \quad F_{n+1}(P,\mathscr{P}^n) = 0. \tag{2.12}$$

In the Euclidean space the tube $\mathscr{T}_n$ of nth order corresponds to the n-dimensional plane containing the points $\mathscr{P}^n$ and the section $\mathscr{S}_{n;P}(\mathscr{T}(\mathscr{P}^n))$ consists of the point P.

The tubes of zeroth and first order are of most interest. One has, for F_1,

$$F_1(P_0,P_1) = 2\sigma(P_0,P_1)$$

and

$$\mathscr{T}_{P_0} = \mathscr{T}(P_0) = \{P \mid \sigma(P_0,P) = 0\}. \tag{2.13}$$

In the proper Euclidean space the tube $\mathscr{T}_{P_0} = \{P_0\}$ consists of one point P_0 and $\mathscr{S}_{0;P_0}(\mathscr{T}_{P_0}) = \{P_0\}$. However, in the pseudo-Euclidean space, which is the space-time in special relativity, $\mathscr{T}_{P_0}$ is the light cone with the vortex at point P_0: its section at point P,

$$\mathscr{S}_{0;P}(\mathscr{T}_{P_0}) = \{P' \mid \sigma(P',P_0) = 0 \wedge \sigma(P_0,P') = \sigma(P_0,P)\} = \mathscr{T}_{P_0}, \tag{2.14}$$

coincides with the light cone.

In describing first-order tubes it is convenient to use the circumstance that $F_2(\mathscr{P}^2)$ can be represented as a product

$$F_2(P_0,P_1,P_2) = S_+(P_0,P_1,P_2)S_2(P_0,P_1,P_2) \times S_2(P_1,P_2,P_0)S_2(P_2,P_0,P_1), \tag{2.15}$$

where

$$S_+(P_0,P_1,P_2) = S(P_0,P_1,) + S(P_1,P_2) + S(P_2,P_0,), \tag{2.16}$$

$$S_2(P_0,P_1,P_2) = S(P_0,P_1,) + S(P_1,P_2) - S(P_0,P_2,). \tag{2.17}$$

Since it follows from Eq. (2.2), S_+ vanishes in only that case if all the terms in Eq. (2.16) vanish. Then no two points form a basis and construction of a tube is not defined. The tube $\mathscr{T}(\mathscr{P}^2)$ can be divided into parts and each of factors (2.17) in Eq. (2.15) is responsible for one part.

Let us set

$$\mathscr{T}_{[P_0P_1]} = \mathscr{T}_{[P_1P_0]} = \{P \mid S_2(P_0,P,P_1) = 0\}, \quad \sigma(P_0,P_1) \neq 0, \tag{2.18}$$

$$\mathscr{T}_{P_0[P_1} = \mathscr{T}_{P_1]P_0} = \{P \mid S_2(P_0,P_1,P) = 0\} \quad \sigma(P_0,P_1) \neq 0. \tag{2.19}$$

One refers to $\mathscr{T}_{[P_0P_1]}$ as the tube segment between the points P_0, P_1 and to $\mathscr{T}_{P_0[P_1}$ as the tube ray outgoing form P_1 in the direction from the point P_0. The set

$$\mathscr{T}_{[P_0P_1} \equiv \mathscr{T}_{[P_0P_1]} \cup \mathscr{T}_{P_0[P_1} \tag{2.20}$$

will be referred to as the tube ray outgoing from P_0 toward point P_1. It is evident from Eq. (2.15) that

$$\mathscr{T}_{P_0P_1} = \mathscr{T}_{P_0]P_1} \cup \mathscr{T}_{[P_0P_1]} \cup \mathscr{T}_{P_0[P_1}. \tag{2.21}$$

Let us use the designations

$$\mathscr{T}_{(P_0P_1)} = \mathscr{T}_{[P_0P_1]} \setminus \left[\bigcup_{l=0}^{l=1} \mathscr{S}_{1;P_l}(\mathscr{T}_{P_0P_1})\right],$$
$$\mathscr{T}_{(P_0P_1]} = \mathscr{T}_{[P_0P_1]} \setminus \mathscr{S}_{1;P_0}(\mathscr{T}_{P_0P_1}), \tag{2.22}$$
$$\mathscr{T}_{P_0)P_1} = \mathscr{T}_{P_0]P_1} \setminus \mathscr{S}_{1;P_0}(\mathscr{T}_{P_0P_1}).$$

Here $\mathscr{T}_{(P_0P_1)}$, $\mathscr{T}_{(P_0P_1]}$ will be referred to as the open and half-open tube intercepts, respectively, between the points P_0, P_1 and $\mathscr{T}_{P_1)P_2}$ will be referred to as the open tube ray outgoing from point P_1 in the direction from the point P_2.

By definition, point P is placed between points P_0 and P_1 if $P \in \mathscr{T}_{(P_0P_1)}$.

Let us use designations (2.7), (2.8), and

$$\mathbf{x} = \mathbf{P}_0\mathbf{P}_1. \tag{2.23}$$

Then in the Euclidean space the first-order tube $\mathscr{T}_{P_0P_1}$ is described by

$$\begin{vmatrix} (\mathbf{e}_1,\mathbf{e}_1) & (\mathbf{e}_1,\mathbf{x}) \\ (\mathbf{x},\mathbf{e}_1) & (\mathbf{x},\mathbf{x}) \end{vmatrix} = 0, \quad (\mathbf{e}_1,\mathbf{e}_1) \neq 0; \tag{2.24}$$

its solution has the form

$$\mathbf{P}_0\mathbf{P} = \mathbf{x} = \mathbf{e}_1\tau + \mathbf{q}, \quad \tau \in \mathbb{R}, \tag{2.25}$$

where $\mathbf{q}$ satisfies the conditions

$$(\mathbf{q},\mathbf{e}_1) = 0, \quad \mathbf{q}^2 = 0. \tag{2.26}$$

In proper Euclidean space Eqs. (2.26) have the unique solution $\mathbf{q} = 0$ and the tube $\mathscr{T}_{P_0P_1}$ coincides with the straight line passing through the points P_0, P_1. In the pseudo-Euclidean space of index 1, metric signature $(+,-,-,...-)$, Eqs. (2.26) have the unique solution $\mathbf{q} = 0$ for the timelike interval P_0P_1, $\mathbf{e}_1^2 = 2\sigma(P_0,P_1) > 0$. For the spacelike interval P_0P_1 $[\sigma(P_0,P_1) < 0]$ there are many solutions of Eqs. (2.26) and $\mathscr{T}_{P_0P_1}$ is not reduced to the straight line $\mathscr{L}_{P_0P_1}$.

In the case of the proper Euclidean space V and in the case of the pseudo-Euclidean space V of index 1 the section $\mathscr{S}_{1;P}(\mathscr{T}_{P_0P_1})$ of the timelike tube $\mathscr{T}_{P_0P_1}$ has the form

$$\mathscr{S}_{1;P}(\mathscr{T}_{P_0P_1}) = \{P\}, \quad P \in \mathscr{T}_{P_0P_1}. \tag{2.27}$$

If the tube is spacelike $[\sigma(P_0,P_1) < 0]$, then

$$\mathscr{S}_{1;Q}(\mathscr{T}_{(P_0P_1)}) = \mathscr{T}_{P_0P_1}(\tau_0), \quad \tau_0 = (\mathbf{P}_0\mathbf{Q},\mathbf{P}_0\mathbf{P}_1)/(\mathbf{P}_0\mathbf{P}_1)^2, \tag{2.28}$$

where $\mathscr{T}_{P_0P_1}(\tau)$ is a set of points P satisfying Eqs. (2.25) and (2.26).

Thus in the four-dimensional event space the timelike tubes are one-dimensional straight lines; however, the spacelike tube $\mathscr{T}_{P_0P_1}$ is a three-dimensional surface formed as a result of moving the light cone section normal to the vector $\mathbf{P_0P_1}$: The section moves in the direction of the vector of $\mathbf{P_0P_1}$.

In the conventional approach a geodesic in a D-dimensional Riemannian space is considered as a *special type of a curve* having extremal properties, as follows.

(i) *Extremality.* The distance $(2\sigma)^{1/2}$ between two points measured along a geodesic is the shortest (extremal) compared to a distance measured along other curves.

(ii) *Definiteness.* Any two points of the geodesic determine unambiguously the geodesic passing through these points.

(iii) *Minimality of section* (one-dimensionality). Any section of a geodesic consists of one point.

In the conventional approach property (ii) is a corollary of property (i) (for the rather small space region); however, property (iii) is a property of any curve (not necessarily of a geodesic).

Another approach is when the geodesic is considered as a *special type of surface* (or of a line tube) that degenerates into a line. Then properties (ii) and (iii) are supposed to be fulfilled; however, property (i) is not defined because the concept of a line (or a curve) is not defined.

Let us try to define a geodesic as a tube having the properties of definiteness and minimality of section at the same time.

Definition 2.5: The tube $\mathscr{T}(\mathscr{P}^n)$ has the definiteness property if for any basis $\mathscr{Q}^n$ of $n+1$ points $\mathscr{Q}^n \subset \mathscr{T}(\mathscr{P}^n)$ the condition

$$\mathscr{T}(\mathscr{Q}^n) = \mathscr{T}(\mathscr{P}^n) \tag{2.29}$$

is fulfilled.

Definition 2.6: The tube $\mathscr{T}(\mathscr{P}^n)$ has the property of section minimality if $\forall P \in \mathscr{T}(\mathscr{P}^n)$,

$$\mathscr{S}_{n;P}(\mathscr{T}(P^n)) = \{P\}. \tag{2.30}$$

Definition 2.7: The σ space is extremal on the tube $\mathscr{T}(\mathscr{P}^n)$ if the conditions of definiteness and section minimality are fulfilled.

Definition 2.8: The σ space is extremal on the set $\mathscr{T}$ of tubes $\mathscr{T}(\mathscr{P}^n)$ of nth order if it is extremal on each tube of the set $\mathscr{T}$.

Definition 2.9: The σ space is extremal in the nth order if it is extremal on all tubes $\mathscr{T}(\mathscr{P}^n)$ of nth order.

Definition 2.10: The tube $\mathscr{T}(\mathscr{P}^n)$ is a geodesic tube $\mathscr{L}(\mathscr{P}^n)$ if the σ space is extremal on the tube $\mathscr{T}(\mathscr{P}^n)$.

Introducing the concept of σ space and considering a geodesic as a kind of a line tube, one hopes to obtain a more adequate description of the event space. The σ space is described in terms of the σ function and only in these terms. In this approach the introduction of such concepts as continuity and manifold is not necessary. In this approach the world lines are replaced by world tubes whose section is a surface, but in general not a point. This approach is associated with the string model, which is popular in the contemporary theory of elementary particles. Finally, quantum particles have no definite world lines: They are "spread" over the event space. The tubes are the world lines spread over the event space. Perhaps the tubes are more adequate for describing quantum particles. In other words, perhaps the σ space describes the real event space at small distances better than does the Riemannian space.

We are not to be discouraged by the circumstance that in the event space the timelike tubes correspond to timelike straight lines and the spacelike tubes correspond to three-dimensional surfaces. Really, world lines of real particles are timelike, whereas there are no spacelike world lines. Perhaps the σ space makes even this distinction.

In reference to the null tubes $\mathscr{T}^*(\mathscr{P}^1)$ $(\sigma(P_0,P_1)=0)$, definition (2.24) or (2.11) does not provide them. For continuous σ space the null tubes $\mathscr{T}^*_{P_0P_1}$ can be defined as follows:

$$\mathscr{T}^*_{P_0P_1} = \lim_{P' \to P_1} \mathscr{T}_{P_0P'}, \quad \sigma(P',P_0) > 0, \quad \sigma(P_0,P_1) = 0 \tag{2.31}$$

as a limit of the timelike tube $\mathscr{T}_{P_0P'}$ at $P' \to P_1$. In the case of pseudo-Euclidean space the result is a null straight line. If the interval P_0P' is spacelike, then the result of Eq. (2.31) depends on the way in which P' is applied to P_1.

Example 1: Let points P of a σ space V be numbered by $n_1, n_2, n_3 \in \mathbb{Z}$, where $\mathbb{Z}$ is the set of all integer numbers. Let the σ function that is between the points $P=(n_1,n_2,n_3)$, $P'=(n_1',n_2',n_3')$, $n_1, n_2, n_3, n_1', n_2', n_3' \in \mathbb{Z}$ be defined by the relation

$$\sigma(P,P') = \frac{a^2}{2} \sum_{i=1}^{3} (n_i - n_i')^2, \quad a = \text{const}, \quad a > 0. \tag{2.32}$$

The σ function depends only on the difference $n_i - n_i'$ $(i=1,2,3)$ and the tube's properties can be investigated without loss of generality in the example of $\mathscr{L}_{P_0P'}$, $P_0=(0,0,0)$. Solving

$$F_2(P_0,P',P) = 0 \tag{2.33}$$

with the σ funciton (2.32) one obtains

$$\mathscr{L}_{P_0P'}: n_i = \tau n_i', \quad i=1,2,3, \quad \tau = k/N, \quad N,k \in \mathbb{Z}, \tag{2.34}$$

where n_i $(i=1,2,3)$ are coordinates of points of $\mathscr{L}_{P_0P'}$, k is an arbitrary integer number, and N is determined by the relation

$$N = \max_{P'' \in \mathscr{L}_{(P_0P')}} \{\sqrt{\sigma(P_0,P')/\sigma(P_0,P'')}\}. \tag{2.35}$$

Thus points of $\mathscr{L}_{P_0P'}$ are points of the straight line $\mathbf{x} = \mathbf{n}'\tau$ $[\mathbf{n}' = (n_1',n_2'n_3'), \mathbf{x} = (x^1,x^2,x^3)]$ with integer coordinates. It is easy to verify that the σ space (2.32) is extremal in the first order.

The detemination of straight lines permits us to introduce the linear vector space over the ring of integer numbers $\mathbb{Z}$. Let us refer to an ordered pair of points P_0, P_1 as the vector $\mathbf{P_0P_1}$. The modulus of the vector $\mathbf{P_0P_1}$ is the number

$$|\mathbf{P_0P_1}| = \sqrt{2\sigma(P_0,P_1)}. \tag{2.36}$$

The vector $\mathbf{P}_0\mathbf{P}' = \alpha\,\mathbf{P}_0\mathbf{P}$ is a result of multiplying the vector $\mathbf{P}_0\mathbf{P}$ by the number $\alpha \in \mathbb{Z}$. The point P' is determined by

$$\sigma(P_0,P') = \alpha^2\sigma(P_0,P), \tag{2.37}$$

with $P' \in \mathscr{L}_{[P_0P}$ if $\alpha \geqslant 0$ and $P' \in \mathscr{L}_{P_0]P}$ if $\alpha < 0$. The sum of two vectors $\mathbf{P}_0\mathbf{P}$ and $\mathbf{P}_0\mathbf{P}'$,

$$\mathbf{P}_0\mathbf{P}'' = \mathbf{P}_0\mathbf{P} + \mathbf{P}_0\mathbf{P}', \tag{2.38}$$

is determined by the relations

$$\begin{aligned} &\mathbf{P}_0\mathbf{P}_1 = 2\mathbf{P}_0\mathbf{P}, \quad \mathbf{P}_0\mathbf{P}_2 = 2\mathbf{P}_0\mathbf{P}', \\ &P'' \in \mathscr{L}_{[P_1P_2]}, \quad S(P'',P_1) = S(P'',P_2). \end{aligned} \tag{2.39}$$

Thus defined operations of summation and multiplication by a number satisfy all the axioms of linear vector space.

The vectors $\mathbf{P}_0\mathbf{P}_1, \mathbf{P}_0\mathbf{P}_2, \ldots, \mathbf{P}_0\mathbf{P}_n$ are referred to as linear independent if the relation

$$\sum_{k=1}^{n} \alpha_k \mathbf{P}_0\mathbf{P}_k = \mathbf{P}_0\mathbf{P}_0 = 0, \quad \alpha_1,\alpha_2,\ldots,\alpha_n \in \mathbb{Z} \tag{2.40}$$

is fulfilled only at $\alpha_i = 0$ $(i = 1,2,\ldots,n)$. The maximal number of the linear-independent vectors is called the dimensionality of the space. In the given case, $D = 3$.

The scalar product of the vectors $\mathbf{P}_0\mathbf{P}$ and $\mathbf{P}_0\mathbf{P}'$ is the number $(\mathbf{P}_0\mathbf{P},\mathbf{P}_0\mathbf{P}')$, which is defined by the relation

$$\begin{aligned} (\mathbf{P}_0\mathbf{P},\mathbf{P}_0\mathbf{P}') &= \Gamma(P_0,P,P') \\ &\equiv \sigma(P_0,P) + \sigma(P_0,P') - \sigma(P,P'). \end{aligned} \tag{2.41}$$

Thus all operations in the vector space are defined only through the world function σ. The space V is a subset of three-dimensional Euclidean space. Indeed, let us imagine that the coordinates (n_1,n_2,n_3), (n_1',n_2',n_3') in expression (2.32) are real numbers. Then relations (2.32)–(2.41) define the three-dimensional proper Euclidean space, where τ in Eq. (2.34) and α in Eq. (2.37) are arbitrary real numbers. The constraint (2.35) is to be omitted.

Example 2: Let points of the σ space V_c be numbered by three numbers n_1, n_2, n_3 $(n_1,n_2,n_3 \in \mathbb{Z},\ 0 \leqslant n_3 < N)$, where N is some natural number and $\mathbb{Z}$ is the set of all integer numbers. Let us define the σ function by the relation

$$\begin{aligned} &\sigma(P,P') = \frac{a^2}{2}\sum_{i=1}^{3}(n_i - n_i')^2, \quad n_1,n_2,n_3,n_1'n_2',n_3' \in \mathbb{Z}, \\ &P = (n_1,n_2,\mathrm{mod}_N n_3), \quad P' = (n_1',n_2',\mathrm{mod}_N n_3'). \end{aligned} \tag{2.42}$$

Here σ is a multivalued function of two points P, P' because the same value $\mathrm{mod}_N(n_3 - n_3')$ corresponds to different values of $n_3 - n_3'$. The space V_c is obtained from the space V of Example 1 by means of identifying those points P whose coordinates n_3 are distinguished by sN, where s is integer. At such an identification $P = P_1 = P_2 = \cdots$ the world function $\sigma(P',P_i)$ $(i = 1,2,\ldots)$ converts into $\sigma(P',P)$ and becomes multivalued, i.e., the world function contains information about the identification. The space V_c is a discrete analog of a cylinder, whereas V from Example 1 is a discrete analog of the three-dimensional plane.

The σ space with the world function (2.42) is not extremal in the first order. In particular, the tube $\mathscr{T}_{P_0P''}$, $P_0 = (0,0,0)$, $P'' = (n_1'',n_2'',\mathrm{mod}_N n_3'')$ consists of geodesics of type (2.34):

$$\begin{aligned} &\mathscr{T}_{P_0P''} = \bigcup_{s\in\mathbb{Z}} \mathscr{L}_{P_0P'}, \quad n_1 = n_1'', \quad n_2'' = n_2', \\ &n_3' = \mathrm{mod}_N n_3'' + sN. \end{aligned}$$

Usually, the difference between a cylinder and a plane is formulated as a distinction of topology of these surfaces. When discussing the distinction between V and V_c, it is hardly appropriate to speak about topology because topology is connected with the concept of continuity, which is not used here. Formally, a "cylindricity" of the space V_c manifests itself in "closed" geodesics consisting of N points. In the σ space V of Example 1 any geodesic contains an infinite number of points.

In Example 2 a single-valued world function can be defined by the relation

$$\begin{aligned} \sigma(P,P') &= (a^2/2)\{(n_1 - n_1')^2 + (n_2 - n_2')^2 \\ &\quad + [\mathrm{mod}_N(n_3 - n_3' + q) - q]_E^2\}, \\ q &= [N/2]_E, \end{aligned} \tag{2.43}$$

where $[\cdots]_E$ denotes the entire part of the number. In this case the space is also no extremal in the first order. Equation (2.43) corresponds to the case when the unique (minimal) value of the function is chosen among many values (2.42). Thus almost all properties of the Euclidean space can be formulated in terms of an extremal σ space without using the concept of continuity. A discrete analog of the Euclidean space can be constructed by removing all points except a countable set, with the values of the world function for the remaining points being conserved.

Definition 2.11: The σ space V, the point set of which is a subset of the points of the Euclidean space E, is a Euclidean σ space.

III. PROPERTIES OF THE σ SPACE

Let us consider a Euclidean space E_n of the dimensionality $n > 1$ and introduce coordinates of an arbitrary point P in the basis $\mathscr{P}^n$ using only the world function. Let $\mathscr{P}^n$ be $n+1$ points that do not lie on one $(n-1)$-dimensional plane. In this case $F_n(\mathscr{P}^n) \neq 0$ and $\mathscr{P}^n$ is a point basis in E_n connected with the basis

$$\mathbf{e}_i = \mathbf{P}_0\mathbf{P}_i, \quad i = 1,2,\ldots,n \tag{3.1}$$

in the linear space of the vectors $\mathbf{P}_0\mathbf{P}$. Then in this basis, according to Eq. (2.7), the metric tensor $g_{ik}(\mathscr{P}^n)$ has form (2.8) and, according to Eq. (2.9),

$$\det\|g_{ik}(\mathscr{P}^n)\| \neq 0, \quad i,k = 1,2,\ldots,n. \tag{3.2}$$

The covariant coordinates x_i of the vector $\mathbf{P}_0\mathbf{P}$ in this basis are

$$\begin{aligned} x_i &= (\mathbf{P}_0\mathbf{P},\mathbf{e}_i) \\ &= \Gamma(P_0,P,P_i) \\ &= \sigma(P_0,P) + \sigma(P_0,P_i) - \sigma(P,P_i), \\ y_i &= \Gamma(P_0,Q,P_i). \end{aligned} \tag{3.3}$$

The world function of two points P,Q of the Euclidean space E_n has the form

$$\sigma(P,Q)=\tfrac{1}{2}g^{ik}(\mathscr{P}^n)(x_i-y_i)(x_k-y_k), \tag{3.4}$$

where g^{ik} are contravariant components of the metric tensor, defined by the relation

$$g^{ik}(\mathscr{P}^n)g_{kl}(\mathscr{P}^n)=\delta_l^i, \quad i,l=1,2,...,n. \tag{3.5}$$

Here the summation is made over repeated superindices and subindices from 1–n.

The following definitions, equivalent to conventional definitions, will be used.

Definition 3.1: The n-dimensional Euclidean space is the set $\mathbb{R}^n$ of all ordered n number $x=\{x_1,x_2,...x_n\}$, where the σ function is given by relations (3.4) and (3.2), $\forall x\in\mathbb{R}$, $\forall y\in\mathbb{R}$.

Definition 3.2: The proper n-dimensional Euclidean space is the Euclidean space for which the equation $\sigma(x,y)=0$, $\forall y\in\mathbb{R}^n$ has the unique solution $x=y$.

Definition 3.3: The pseudo-Euclidean space is the Euclidean space, which is not proper.

Definition 3.4: The flat space is the Riemannian space with an everywhere vanishing Riemannian curvature tensor. A flat space can differ from the Euclidean space in topology.

Let us consider a $(n+2)\times(n+2)$ matrix

$$\mathscr{A}_{n+2}=\|a_{ik}\|, \quad a_{ik}=\Gamma(P_0,P_i,P_k),$$
$$i,k=1,2,...,n+2, \quad P_{n+1}=P, \quad P_{n+2}=Q, \tag{3.6}$$
$$F_n(\mathscr{P}^n)\neq 0. \tag{3.7}$$

Theorem 3.1: Let $\mathscr{P}^n$ be a basis of $n+1$ points in the σ space V, i.e., Eq. (3.7) is fulfilled. Then for the tube $\mathscr{T}(\mathscr{P}^n)$ to be a Euclidean σ space, fulfillment of the following conditions is necessary and sufficient:

(i) The rank $(\mathscr{A}_{n+2})=n$, $\forall P\in\mathscr{T}(\mathscr{P}^n)$, $Q\in\mathscr{T}(\mathscr{P}^n)$.

(ii) The section $\mathscr{T}(\mathscr{P}^n)$ is minimal at any point: $\mathscr{S}_{n;Q}(\mathscr{T}(\mathscr{P}^n))=\{Q\}$, $Q\in\mathscr{T}(\mathscr{P}^n)$.

Proof: Let $\mathscr{T}(\mathscr{P}^n)\subset E_n$ be a subset of a Euclidean space E_n and $P\in\mathscr{T}(\mathscr{P}^n)$, $Q\in\mathscr{T}(\mathscr{P}^n)$. Then using the designations (2.8), (3.3), and (3.7) and property (3.4), one can represent the matrix in the form

$$\mathscr{A}_{n+2}=\begin{array}{c|c|c} g_{ik}(\mathscr{P}^n) & x_i & y_i \\ \hline x_k & \sum_{l=1}^{n} x_l y^l & \sum_{l=1}^{n} y_l x^l \\ \hline y_k & \sum_{l=1}^{n} x_l y^l & \sum_{l=1}^{n} y_l y^l \end{array}, \quad i,k=1,2,...,n, \tag{3.8}$$

where

$$x^i=\sum_{k=1}^{n} g^{ik}(\mathscr{P}^n)x_k, \quad y^i=\sum_{k=1}^{n} g^{ik}(\mathscr{P}^n)y_k \tag{3.9}$$

are contravariant coordinates of points P and Q in the same basis $\mathscr{P}^n$.

According to Eq. (3.7),

$$g\equiv\det\|g_{ik}\|=F_n(\mathscr{P}^n)\neq 0 \tag{3.10}$$

and rank r of the matrix $\mathscr{A}_{n+2}$ is not less than n ($r\geqslant n$). On the other hand, the last and next to last columns of the matrix (3.8) are linear combinations of the first n columns with the factors y^k and x^k ($k=1,2,...,n$), respectively, i.e., rank of the matrix $r\leqslant n$. Then one concludes that in the Euclidean space the rank of the matrix $\mathscr{A}_{n+2}$ is equal to n provided that when fulfilling Eq. (3.7),

$$\text{rank}\,(\mathscr{A}_{n+2})=n. \tag{3.11}$$

It is easy to verify that the section of any plane $\mathscr{T}(\mathscr{P}^n)$ of the n-dimensional Euclidean space is minimal in the sense of Eq. (2.30).

Now let $n+1$ points $\mathscr{P}^n$ of a σ space V satisfy Eq. (3.7). Let $P\in\mathscr{T}(\mathscr{P}^n)$, $Q\in\mathscr{T}(\mathscr{P}^n)$, rank of matrix (3.6), be equal to n and any section of the tube $\mathscr{T}(\mathscr{P}^n)$ be minimal, i.e., Eq. (2.30) is fulfilled $\forall P\in\mathscr{T}(\mathscr{P}^n)$. According to Eq. (2.11), the conditions $P\in\mathscr{T}(\mathscr{P}^n)$, $Q\in\mathscr{T}(\mathscr{P}^n)$ mean that

$$M_{n+1,n+1}=F_{n+1}(P,\mathscr{P}^n)=\begin{vmatrix} g_{ik} & x_i \\ x_k & 2\sigma(P_0P) \end{vmatrix}=0,$$
$$M_{n+2,n+2}=F_{n+1}(Q,\mathscr{P}^n)=0, \tag{3.12}$$

where $M_{n+1,n+1}$ and $M_{n+2,n+2}$ are two principal minors of the matrix (3.6). It follows from Eq. (3.12) that

$$\sigma(P_0,P)=\tfrac{1}{2}\sum_{i,k=1}^{n} g^{ik}x_i x_k, \quad \forall P\in\mathscr{T}(\mathscr{P}^n). \tag{3.13}$$

Using Eq. (3.13), the matrix $\mathscr{A}_{n+2}$ can be written in the form

$$\mathscr{A}_{n+2}=\begin{array}{c|c|c} g_{ik} & x_i & y_i \\ \hline x_k & \sum_{l=1}^{n} x_l y^l & \Gamma(P_0,P,Q) \\ \hline y_k & \Gamma(P_0,P,Q) & \sum_{l=1}^{n} y_l y^l \end{array}, \tag{3.14}$$

In order that the rank of matrix (3.14) be equal to n, it is necessary that each of the two last columns are a linear combination of the first n columns: This means that

$$\Gamma(P_0,P,Q)=\sum_{j=1}^{n} x_j y^j. \tag{3.15}$$

As a result of definition (3.3) and Eq. (3.13), (3.15) leads to Eq. (3.4).

Now let us map $\mathscr{T}(\mathscr{P}^n)\to\Omega_n\subset\mathbb{R}^n$, where $\mathbb{R}^n$ is a set of all elements consisting of n ordered real numbers $x=\{x_i\}$ ($i=1,2...,n$). Let us correspond to each point P its covariant coordinates in the basis $(\mathscr{P}^n)$. Such a mapping is one-to-one as a result of the minimal section condition (2.30), i.e., each point $P\in\mathscr{T}(\mathscr{P}^n)$ corresponds to one point $x\in\Omega_n\subset\mathbb{R}^n$ and each point $x\in\Omega_n$ corresponds to only one point $P\in\mathscr{T}(\mathscr{P}^n)$. Really, if the points $P\in\mathscr{T}(\mathscr{P}^n)$ and $Q\in\mathscr{T}(\mathscr{P}^n)$ have the same coordinates

$$x_i=\Gamma(P_0,P_i,P)=\Gamma(P_0,P_i,Q)=y_i, \quad i=1,2,...,n, \tag{3.16}$$

then it follows from Eq. (3.13) that

$$\sigma(P_0,P)=\sigma(P_0,Q). \tag{3.17}$$

As a corollary of Eqs. (2.3) and (3.16) one has

$$\sigma(P_i,P)=\sigma(P_i,Q), \quad i=1,2,...,n. \tag{3.18}$$

Conversely, Eq. (3.16) follows from Eqs. (3.17), (3.18), and (2.3). Then definition (2.12) of the tube $\mathscr{T}(\mathscr{P}^n)$ section can also be represented in the form

$$\mathscr{S}_{n;P}(\mathscr{T}(\mathscr{P}^n)) = \Big\{P' | F_{n+1}(P', \mathscr{P}^n) = 0 \underset{l=1}{\overset{l=n}{\wedge}} \Gamma(P_0, P_l, P') = \Gamma(P_0, P_i, P)\Big\},$$

$$F_{n+1}(P, \mathscr{P}^n) = 0. \tag{3.19}$$

It follows from (3.19) that if $Q \in \mathscr{T}(\mathscr{P}^n)$ and has the coordinates (3.16), then $Q \in \mathscr{S}_{n;P}(\mathscr{P}^n)$. As a result of the minimal section condition the section consists of one point. Then any coordinates x correspond to the unique point P and the mapping $\mathscr{T}(\mathscr{P}^n) \to \Omega_n$ is one-to-one.

Thus for $x \in \Omega_n \subset \mathbb{R}^n$, $y \in \Omega_n \subset \mathbb{R}^n$ the σ function has the form (3.4). Let us propagate Eq. (3.4) over all the points of the set $\mathbb{R}^n$. Then the $\mathbb{R}^n$ with the σ function defined on it is a Euclidean space. The Ω_n with the σ function defined on it is a subset of $\mathbb{R}^n$. Hence, Ω_n is the Euclidean σ space. As a result of the one-to-one correspondence $\mathscr{T}(\mathscr{P}^n) \leftrightarrow \Omega_n$ and the invariance of the σ function one concludes that $\mathscr{T}(\mathscr{P}^n)$ is the Euclidean σ space. Theorem (3.1) is proved.

Theorem (3.1) permits us to determine the dimensionality D of the proper Euclidean σ space V based only upon the σ function. Let us use the following procedure. Let $P_0 \in V$. If $\mathscr{T}_{P_0} = V$, then $D = 0$; otherwise, $\exists P_1 \notin \mathscr{T}_{P_0}$, $P_1 \in V$. Then $F_1(\mathscr{P}^1) \neq 0$, $D > 0$. If $\mathscr{T}(\mathscr{P}^1) = V$, $\mathscr{P}^1 = \{P_0, P_1\}$, then $D = 1$; otherwise $\exists P_2 \notin \mathscr{T}(\mathscr{P}^1)$, $P_2 \in V$, and $D > 1$. If $\mathscr{T}(\mathscr{P}^2) = V$, then $D = 2$; otherwise $\exists P_3 \notin \mathscr{T}(\mathscr{P}^2)$, $P_3 \in V$, and $D > 2$, etc. Let us continue the procedure until at some n, $\mathscr{T}(\mathscr{P}^n) = V$ and $D = n$. Such a procedure can be produced in any σ space, but for the single-valued result it is necessary that it does not depend on the choice of the basis $\mathscr{P}^n$: It is provided if V is a proper Euclidean σ space.

IV. THE EUCLIDEAN σ SPACE AND EXTREMALITY

The proper Euclidean space is extremal in all orders. Let us investigate whether the reverse statement is valid: A σ space that is extremal in all orders is proper Euclidean. In general, this statement is not valid, although its violations are rather an exception than a rule.

Let a basis $\mathscr{P}^n \subset \Omega$, where Ω is the set of points of the extremal σ space V, $P \in \mathscr{L}(\mathscr{P}^n)$, and $Q \in \mathscr{L}(\mathscr{P}^n)$. The definiteness condition (2.29) containing $n+3$ points P, Q, $\mathscr{P}^n$ can be written in the form

$$F_{n+1}(P, \mathscr{P}^n) = 0, \quad F_{n+1}(Q, \mathscr{P}^n) = 0, \tag{4.1}$$

$$F_{n+1}(P, Q, \mathscr{P}^n_l) = 0, \quad l = 0,1,...,n. \tag{4.2}$$

Here the $n+3$ equations (4.1) and (4.2) (except for the case $l = 0$) represent the condition of vanishing $n+2$ principal minors of the matrix (3.6). Using the designations (3.2) and (3.3), one obtains from Eq. (4.1) relation (3.13) and a similar relation with the substitution $P \to Q$, $x \to y$. Each of Eqs. (4.2) represents a quadratic equation with respect to the variable

$$z = \Gamma(P_0, P, Q) - \sum_{i=1}^{n} x_i y^i. \tag{4.3}$$

Equations (4.2) have the trivial solution

$$z_1 = 0, \tag{4.4}$$

which corresponds to relation (3.15) and leads to expression (3.4) for $\sigma(P,Q)$, i.e., to the Euclidean σ space $\mathscr{T}(\mathscr{P}^n)$.

Using the trivial solution (4.4), one can reduce the order of Eqs. (4.2). Then Eqs. (4.2) (except for the case $l = 0$) are reduced to n linear equations with respect to z:

$$g^{ll} z + 2x^l y^l = 0, \quad l = 1,2,...,n. \tag{4.5}$$

One can show that

$$(x^l)^2 = \frac{F_n(P, \mathscr{P}^n_l)}{F_n(\mathscr{P}^n)}, \quad (y^l)^2 = \frac{F_n(Q, \mathscr{P}^n_l)}{F_n(\mathscr{P}^n)},$$
$$l = 1,2,...,n, \quad g^{ll} = \frac{F_{n-1}(\mathscr{P}^n_l)}{F_n(\mathscr{P}^n)}, \quad l = 1,2,...,n. \tag{4.6}$$

Let us substitute Eqs. (4.6) into Eqs. (4.5) and take into account that the first ($l = 0$) of Eqs. (4.2) is obtained from $l = k$ as a result of substituting $P_0 \leftrightarrow P_k$. Then one obtains a system of n linear equations for z:

$$F_{n-1}(\mathscr{P}^n_l) z + 2\sqrt{F_n(P, \mathscr{P}^n_l) F_n(Q, \mathscr{P}^n_l)} = 0, \quad l = 0,1,...,n. \tag{4.7}$$

The coefficients of Eqs. (4.7) do not depend on $\sigma(P,Q)$: The condition of a common solution of Eqs. (4.7) imposes constraints upon these coefficients, which are constraints upon coordinates of the points P and Q.

Only the nontrivial solution of system (4.7) is of interest. The trivial solution $z = 0$ returns us to case (4.4) of the Euclidean σ space.

If the points $\mathscr{P}^n = \{P_0, P_1, ..., P_n\}$ are such that no n lie on a tube of $(n-2)$th order, i.e.,

$$F_{n-1}(\mathscr{P}^n_l) \neq 0, \tag{4.8}$$

then according to Eq. (4.6),

$$g^{ll} \neq 0, \quad l = 1,2,...,n \tag{4.9}$$

and Eq. (4.5) can be represented in the form

$$\frac{x^l y^l}{g^{ll}} = \frac{\sqrt{F_n(P, {}^l\mathscr{P}^n) F_n(Q, {}^l\mathscr{P}^n)}}{F_{n-1}({}^l\mathscr{P}^n)}, \quad l = 1,2,...,n. \tag{4.10}$$

Definition 4.1: The $(n+1)$ point basis $\mathscr{P}^n$ in σ space V is the nondegenerate basis provided that any n points of $\mathscr{P}^n$ form an n-point basis $\mathscr{P}^n_l$ $(F_{n-1}(\mathscr{P}^n_l) \neq 0, l = 0,1,...,n)$.

Definition 4.2: Points P and Q of the σ space V are mutually conjugate with respect to the nondegenerate $(n+1)$ point basis $\mathscr{P}^n \subset V$ if $P \notin \mathscr{T}(\mathscr{P}^n_l)$, $Q \notin \mathscr{T}(\mathscr{P}^n_l)$ $(l = 0,1,...n)$, and their contravariant coordinates x and y satisfy Eq. (4.10).

If one of coordinates $x^l = 0$ [which is equivalent to $P \in \mathscr{T}(\mathscr{P}^n_l)$ or $F_n(P, \mathscr{P}^n_l) = 0$], then according to Eqs. (4.5)–(4.10) one obtains the trivial solution (4.4).

Fixing the coordinates $x^i \neq 0$ $(i = 1,2,...,n)$ of point P, one concludes from Eq. (4.10) that the set of points Q satisfying Eq. (4.10) with an indefinite rhs forms a straight line passing through point P_0. From the symmetry under consideration one concludes that point Q must lie on the straight line passing through point P_1, through P_2,... . Thus in this case there is no more than one point Q conjugate to P with respect to the basis $\mathscr{P}^n$.

One can formulate the following theorems.

Theorem 4.1: Let $\mathscr{P}^n$ be a nondegenerate $(n+1)$ point basis in the σ space V, which is extremal on $\mathscr{T}(\mathscr{P}^n) = \mathscr{L}(\mathscr{P}^n)$. If $P \in \mathscr{L}(\mathscr{P}^n)$, $P \notin \mathscr{T}(\mathscr{P}_l^n)$ $(l = 0,1,...,n)$, then there exists no more than one point Q conjugate to P with respect to $\mathscr{P}^n$.

Theorem 4.2: Let $\mathscr{P}^n$ form a nondegenerate $(n+1)$ point basis in the σ space V and V be extremal on $\mathscr{T}(\mathscr{P}^n) = \mathscr{L}(\mathscr{P}^n)$. If $Q, P \in \mathscr{L}(\mathscr{P}^n)$, $P, Q \notin \mathscr{L}(\mathscr{P}_l^n)$ $(l = 0,1,...,n)$, then either the σ space $\mathscr{L}(\mathscr{P}^n)$ is Euclidean or the points P and Q are mutually conjugate with respect to $\mathscr{P}^n$.

Theorem 4.3: Let $\mathscr{P}^n$ be a nondegenerate $(n+1)$ point basis in the σ space V and V be extremal on $\mathscr{T}(\mathscr{P}^n) = \mathscr{L}(\mathscr{P}^n)$. Then the σ space $V_1 = \mathscr{L}(\mathscr{P}^n)$ is Euclidean provided that the number N of points in $V_1 = \mathscr{L}(\mathscr{P}^n)$ is distinguished from $n+3$.

Proof: Insofar as $\mathscr{P}^n \subset \mathscr{L}(\mathscr{P}^n)$, then $N \geqslant n+1$. If $N = n+1$, then $V_1 = \mathscr{P}^n$ and Theorem 4.3 is evident from Eqs. (3.3), (3.5), and (2.8).

If $N = n+2$, $V_1 = \{\mathscr{P}^n, P\}$. Then it follows from the first of Eqs. (3.12) that relation (3.13) takes place. Thereafter, one can verify Eq. (3.4) using Eqs. (3.3), (2.8), (3.5), and (3.13).

Now, if $N = n+4$, then $V_1 = \{\mathscr{P}^n, P, Q, R\}$. Let V_1 be non-Euclidean. Then according to Theorem 4.2 Q is conjugate to P and R is conjugate to P with respect to the basis $\mathscr{P}^n$. According to Theorem 4.1 the conjugate point is unique; then $Q \neq R$ cannot exist. Then the σ space is Euclidean. The same consideration can be given in the case of $N > n+4$. Theorem 4.3 has been proved.

Now let us consider an illustration for Theorem 4.3: When $N = n+3$, the σ space is extremal on $\mathscr{T}(\mathscr{P}^n)$; however, $\mathscr{T}(\mathscr{P}^n)$ is non-Euclidean.

Example 3: Let the σ space Ω_4 consist of the four points 0, 1, P, Q.

$$\begin{aligned} &S(0,1) = S(P,Q) = a, \\ &S(0,P) = S(Q,1) = b, \quad 0 < b < a. \\ &S(P,1) = S(0,Q) = a - b, \end{aligned} \tag{4.11}$$

It is easy to verify that

$$\mathscr{T}_{01} = \mathscr{T}_{0P} = \mathscr{T}_{0Q} = \mathscr{T}_{1P} = \mathscr{T}_{1Q} = \mathscr{T}_{PQ} = \Omega_4,$$

i.e., the tube $\mathscr{T}_{01}$ contains all points of the σ space Ω_4 and Ω_4 is one-dimensional. The line tube has the definiteness property. Besides, if $a - b \neq b$, then the section of the tube is minimal. Thus at $a \neq 2b$ the σ space is extremal in the first order.

In addition, one has the following situation, which is unusual for the proper Euclidean space:

$$P \in \mathscr{L}_{(01)}, \quad Q \in \mathscr{L}_{(01)},$$

i.e., the points P and Q are placed between the points 0 and 1. At the same time, $0 \in \mathscr{L}_{(PQ)}$, $1 \in \mathscr{L}_{(PQ)}$, i.e., the points 0 and 1 lie between the points P and Q. Here Ω_4 is not a Euclidean σ space.

Relations (4.11) can be understood from the Euclidean point of view if one imagines that the points are placed on a closed geodesic in the order $0,P,1,Q,0,...$. For the transition to the unclosed geodesic it is sufficient to substitute $S(P,Q) = a$ with $S(P,Q) = a - 2b$.

Another version of the interpretation is shown in Fig. 1. One-dimensional proper Euclidean σ space consists of the five points Q_1', 0, P, 1, Q_0'. The points Q_0' and Q_1' are mirror images of point Q at a reflection with respect to points 1 and 0, respectively. The point Q is conjugate to point P with respect to the basis (0,1) provided that $S(P,Q_1') = S(P,Q_0')$. Let the points P and Q be conjugate with respect to the basis (0,1). Then it is possible to identify the points Q_1' and Q_0', denoting them by means of Q and conserving all distances except $S(Q_0',Q_1')$, $S(0,Q_0')$, and $S(1,Q_1')$. As a result the σ space (4.11) arises: It is flat, but non-Euclidean.

Example 4: Let the σ space Ω_5 consist of the five points 0, 1, 2, P, Q.

$$\begin{aligned} &S(0,1) = S(0,2) = S(1,2) = a, \\ &S(0,P) = a/\sqrt{3} + \varepsilon + O(\varepsilon^2), \\ &S(1,P) = S(2,P) = a/\sqrt{3} - \varepsilon/2 + O(\varepsilon^2), \\ &S(1,Q) = S(2,Q) = a/\sqrt{3} + \varepsilon/2 + O(\varepsilon^2), \\ &S(0,Q) = a/\sqrt{3} - \varepsilon + O(\varepsilon^2), \\ &S(P,Q) = a/\sqrt{3} + O(\varepsilon^2), \quad \varepsilon \ll 1. \end{aligned} \tag{4.12}$$

One can verify that $\mathscr{T}_{012} = \Omega_5$ and Ω_5 is a two-dimensional σ space extremal in the second order. However, Ω_5 is not a Euclidean σ space. The disposition of points on the proper Euclidean plane is shown in Fig. 2. All distances (4.12) except $S(P,Q)$ correspond to Euclidean distances in Fig. 2. If the distance $S(P,Q)$ were Euclidean, then one would have $S(P,Q) = 2\varepsilon + O(\varepsilon^2)$.

The properties of Ω_5 can be understood if one considers the σ spaces $\Omega_5 \setminus \{0\}$, $\Omega_5 \setminus \{1\}$, $\Omega_5 \setminus \{2\}$, $\Omega_5 \setminus \{P\}$, and $\Omega_5 \setminus \{Q\}$ consisting of four points. Each of them is a Euclidean σ space. The disposition of points in these spaces is shown in Fig. 2. The point Q in the first three σ spaces is replaced, respectively, by Q_0', Q_1', and Q_2' in the picture of σ spaces on the proper Euclidean plane. Thus one has $\Omega_5 \setminus \{0\} = \{1,2,P,Q_0'\}$, $\Omega_5 \setminus \{1\} = \{0,2,P,Q_1'\}$, $\Omega_5 \setminus \{2\} = \{0,1,P,Q_2'\}$, $\Omega_5 \setminus \{P\} = \{0,1,2,Q\}$, and $\Omega_5 \setminus \{Q\} = \{0,1,2,P\}$. Then all the distances shown in Fig. 2 coincide with those in Eq. (4.12). The presented examples illustrate Theorem 4.3 in the case when $N = n+3$ and σ space is non-Euclidean.

V. THE DENSE σ SPACE AND CURTAILED TUBES

Definition 5.1: A non-null tube ray $\mathscr{T}_{[P_0,P_1}$ $(\sigma(P_0,P_1) \neq 0)$ is dense at point P_0 if an infinite-converging-to-P_0 sequence $P_1', P_2', ...$ of distinguishing in pairs points can be found on the open tube ray $\mathscr{T}_{(P_0,P_1}$. This means that for $\forall \varepsilon > 0$ there exists N_ε such that $|2\sigma(P_0,P_n')|^{1/2} < \varepsilon$ if $n > N_\varepsilon$.

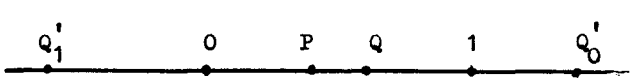

FIG. 1. Version of interpretation.

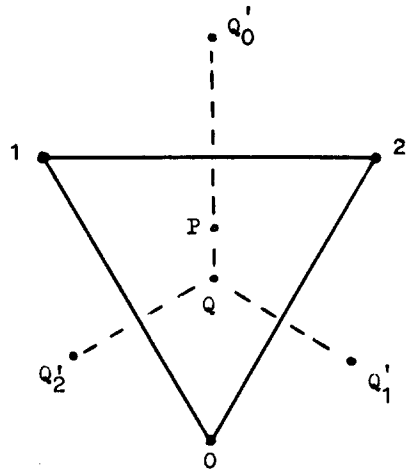

FIG. 2. Disposition of points on the proper Euclidean plane.

Definition 5.2: The σ space V is dense at point P_0 in the timelike direction if any timelike tube ray $\mathscr{T}_{[P_0P}$ is dense at P_0.

Definition 5.3: The σ space V is dense at point P_0 if any non-null tube ray $\mathscr{T}_{[P_0P}$ is dense at P_0.

Definition 5.4: If a non-null tube ray $\mathscr{T}_{[P_0Q} \subset V$ is dense at the point P_0 and the limit

$$(\mathbf{u}_{P_0Q},\mathbf{P_0P}) = \left|\sqrt{2\sigma(P_0,Q)}\right| \Pi\,(P_0,Q,P) = \lim_{\substack{Q'\to P_0 \\ Q'\in\mathscr{T}_{(P_0Q}}} \left|\sqrt{\frac{\sigma(P_0,Q)}{\sigma(P_0,Q')}}\right| \Gamma(P_0,Q',P) \tag{5.1}$$

exists $\forall P\in V$ and for any way of tending Q' to P_0, then the tube ray $\mathscr{T}_{[P_0Q}$ determines a non-null direction vector $\mathbf{u}_{P_0Q}$ at point P_0. Here $\Pi(P_0,Q,P)$ is a projection of the vector $\mathbf{P_0P}$ onto the direction $\mathbf{d}_{P_0Q}$.

The limit (5.1) is defined in the conventional way. For any $\varepsilon>0$ such $\delta_\varepsilon>0$ exists, that the inequality

$$|\Pi\,(P_0,Q,P) - \Gamma(P_0,Q',P)|2\sigma(P_0,Q')|^{-1/2}| < \varepsilon \tag{5.2}$$

is a consequence of the conditions

$$\left|\sqrt{2\sigma(P_0,Q')}\right| < \delta, \quad Q'\in\mathscr{T}_{(P_0,Q}. \tag{5.3}$$

Definition 5.4 can be formulated briefly as follows: $\mathbf{u}_{P_0Q} = \mathrm{gen}(\mathscr{T}_{[P_0Q})$ ("$\mathbf{u}_{P_0Q}$ is generated by $\mathscr{T}_{[P_0Q}$").

Definition 5.5: If the σ space V is dense at the point $P_0\in V$ and if the equation

$$\frac{(\mathbf{u}_{P_0Q'},\mathbf{P_0P})}{\left|\sqrt{2\sigma(P_0,Q')}\right|} = \frac{(\mathbf{u}_{P_0,Q},\mathbf{P_0P})}{\left|\sqrt{2\sigma(P_0,Q)}\right|}, \quad \sigma(P_0,Q)\neq 0, \quad \forall P\in V \tag{5.4}$$

considered as an equation for the determination of Q', has only the solutions $Q'\in\mathscr{T}_{[P_0Q}$, then the non-null direction vector $\mathbf{u}_{P_0Q}$ at point P_0 determines the non-null tube ray $\mathscr{T}_{[P_0Q}$.

Definition 5.5. is formulated briefly as follows: $\mathscr{T}_{[P_0Q} = \mathrm{gen}(\mathbf{u}_{P_0Q})$ ($\mathscr{T}_{[P_0Q}$ is generated by $\mathbf{u}_{P_0Q}$). This property is a variety of the definiteness property (Definition 2.5) when the two points P_0 and Q' defining the tube are infinitely close.

In the conventional approach the concept of direction is connected with the concept of a curve. The latter is connected with the possibility of one-to-one continuous mapping into the set $\mathbf{R}$ of all real numbers. In our approach the concept of direction is connected with the extremality of $\mathscr{T}_{P_0Q}$ in Eq. (5.1) because only in this case does the limit exist as a rule. It is quite natural since the concepts of a curve and an extremal tube that is dense at all points are rather close.

However, there are σ spaces that do not contain the extremal tubes $\mathscr{T}_{P_0P}$. For instance, the D-dimensional pseudo-Euclidean space E_D of index i ($g_{kl} = \mathrm{diag}(1,1,...,1,-1,-1,...,-1)$, $1<i<D-1$, $D\geqslant 4$) is not extremal on any tube $\mathscr{T}_{P_0P}$. In this case nonextremality in the first order is caused by nonextremality in the zeroth order. The extremal tube ray $\mathscr{T}_{P_0P}$ can be obtained from the nonextremal tube ray by means of the special curtailing operation $\mathscr{C}$:

$$\mathscr{T}^c_{[P_0P} = \mathscr{C}\,\mathscr{T}_{[P_0P}. \tag{5.5}$$

There are different ways of defining the curtailing operation. One way is the operation of intense definiteness (intense extremality)

$$\mathscr{T}^c_{[P_0P} = \mathscr{I}\,\mathscr{T}_{[P_0P} = \{P'|\mathscr{T}_{[P_0P'} = \mathscr{T}_{[P_0P}\} \tag{5.6}$$

when the set $\mathscr{T}^c_{[P_0P}$ contains only those points P' of $\mathscr{T}_{[P_0P}$ for which $\mathscr{T}_{[P_0P'}$ coincides with $\mathscr{T}_{[P_0P}$. Here the σ space is extremal on $\mathscr{T}_{[P_0P_1}$ if

$$\mathscr{T}^c_{[P_0P'} = \mathscr{T}^c_{[P_0P_1}, \quad P'\in\mathscr{T}^c_{[P_0P_1}, \quad P'\neq P_0, \tag{5.7}$$

$$\mathscr{S}_{1;Q}(\mathscr{T}^c_{[P_0P_1}) = \{P'|P'\in\mathscr{T}^c_{[P_0P_1} \bigwedge_{l=0}^{l=1} \sigma(P_l,Q) = \sigma(P_l,P')\} = \{Q\}, \quad Q\in\mathscr{T}^c_{[P_0P_1}. \tag{5.8}$$

Example 5: Let us consider the pseudo-Euclidean space E_4 of index 2. In the Cartesian coordinates $x = (x^1,x^2,x^3,x^4)$ the σ function has form (3.4), with $g_{ik} = \mathrm{diag}(1,1,-1,-1)$. Let $P_0 = (0,0,0,0)$, $P_1 = (1,0,0,0)$, and $P_2 = (0,0,1,0)$. Let $P = (x^1,x^2,x^3,x^4)$ be a running point of the set. The calculation leads to

$$\mathscr{T}_{[P_0P_1}:(x^2)^2 - (x^3)^2 - (x^4)^2 = 0 \wedge x^1\geqslant 0,$$
$$\mathscr{T}^c_{[P_0P_1} = \mathscr{I}\,\mathscr{T}_{[P_0P_1}:x^2 = x^3 = x^4 = 0 \wedge x^1\geqslant 0,$$
$$\mathscr{T}_{[P_0P_2}:(x^1)^2 + (x^2)^2 - (x^4)^2 = 0 \wedge x^3\geqslant 0,$$
$$\mathscr{T}^c_{[P_0P_1} = \mathscr{I}\,\mathscr{T}_{[P_0P_1}:x^1 = x^2 = x^4 = 0 \wedge x^3\geqslant 0.$$

Thus the curtailing operation transforms the three-dimensional tube ray $\mathscr{T}_{[P_0P_1}$ into the one-dimensional curtailed tube ray $\mathscr{T}^c_{[P_0P_1}$, where the σ space is extremal on $\mathscr{T}^c_{[P_0P_1}$ in the sense of Eqs. (5.7) and (5.8), where Eq. (5.7) is fulfilled as a result of Eq. (5.6).

Further, the σ space that is extremal on all timelike tubes $\mathscr{T}_{P_0P}$ will be considered. Spacelike tubes are not in general extremal. Such a σ space is important from a physical viewpoint because the real event space is a σ space of this kind.

Definition 5.6: The scalar product between two non-null direction vectors $\mathbf{u}_{P_0P}$ and $\mathbf{u}_{P_0Q}$ at P_0 (or between two non-null tube rays $\mathscr{T}_{[P_0P}$ and $\mathscr{T}_{[P_0Q}$) is determined if the following limit exists:

$$(\mathbf{u}_{P_0P},\mathbf{u}_{P_0Q}) \equiv \Gamma^*_{P_0}(P_0,P,Q) = \lim_{Q'\to P_0,Q'\in\mathscr{T}_{(P_0Q}} \lim_{P'\to P_0,P'\in\mathscr{T}_{(P_0P}} \left|\sqrt{\frac{\sigma(P_0,P)\sigma(P_0,Q)}{\sigma(P_0,P')\sigma(P_0,Q')}}\right| \times \Gamma(P_0,P',Q'). \tag{5.9}$$

Let a σ space V be given on the set Ω. On the set

$$\Omega^+_{P_0} = \{P\,|\sigma(P_0,P)>0\}\cup\{P_0\} \tag{5.10}$$

there are defined two kinds of objects: the vector $\mathbf{P_0P}$ (the point $P\in\Omega^+_{P_0}$ determines $\mathbf{P_0P}$) and the direction vector $\mathbf{u}_{P_0P}$, $P\in\Omega^+_{P_0}$.

The scalar products between the two vectors are defined by Eqs. (5.1), (5.9), and

$$(\mathbf{P_0P},\mathbf{P_0Q}) = \Gamma(P_0,P,Q). \tag{5.11}$$

Two different σ spaces can be defined on $\Omega^+_{P_0}$: the σ space $V^+_{P_0}$ with the world function σ and the σ space $V^{+*}_{P_0}$ with the world function

$$\sigma^*_{P_0}(P,Q) = \sigma(P_0,P)+\sigma(P_0,Q) - \Gamma^*_{P_0}(P_0,P,Q), \quad P,Q\in\Omega^+_{P_0}. \tag{5.12}$$

It is supposed that the vectors $\mathbf{u}_{P_0P}$ ($P\in\Omega^+_{P_0}$) belong to $V^{+*}_{P_0}$, whereas the vectors $\mathbf{P_0P}$ ($P\in\Omega^+_{P_0}$) belong to $V^+_{P_0}$.

Definition 5.7: The σ space V given on a set Ω determines a σ space $V^{+*}_{P_0}$, with the σ function defined by Eq. (5.12) on the set $\Omega^+_{P_0}$ if the following conditions are fulfilled.

(i) The σ space V is dense at P_0 and extremal on any timelike tube ray $\mathscr{T}_{[P_0Q}\ni P_0$, $(\sigma(P_0,Q)>0)$, $Q\in\Omega$.

(ii) Any timelike tube ray $\mathscr{T}_{[P_0Q}$ $(\sigma(P_0,Q)>0)$ determines the timelike direction vector $\mathbf{u}_{P_0Q}$ $(\mathbf{u}_{P_0Q}=\mathrm{gen}(\mathscr{T}_{[P_0Q}), \sigma(P_0,Q)>0)$.

(iii) The scalar product of two direction vectors $\mathbf{u}_{P_0P}$ and $\mathbf{u}_{P_0Q}$ $(\sigma P_0,Q)>0$, $\sigma(P_0,P)>0)$ at P_0 is determined.

It is easy to verify that the function $\Gamma^*_{P_0}(P_0,P,Q)$ calculated by means of Eq. (2.3) with the σ function (5.12) coincides with expression (5.9).

Definition 5.8: The direction vector $\mathbf{u}_{P_0P}$ ($P\in\Omega$) is tangent to the vector $\mathbf{P_0P}$ ($P\in\Omega$) on $\Omega^+_{P_0}$ if

$$(\mathbf{u}_{P_0Q},\mathbf{P_0P}) = (\mathbf{u}_{P_0Q},\mathbf{u}_{P_0P}) = \Gamma^*_{P_0}(P_0,P,Q), \quad \forall Q\in\Omega^+_{P_0}. \tag{5.13}$$

Definition 5.9: The σ space V given on the set Ω determines the D-dimensional σ space $V^+_{P_0Dt}$ of the timelike tangent direction vectors $\mathbf{u}_{P_0Q}$ on the set $\Omega^+_{P_0}$ if the following conditions are fulfilled.

(i) Here V determines the σ space $V^{+*}_{P_0}$ with the σ function (5.12) on $\Omega^+_{P_0}$.

(ii) Any direction vector $\mathbf{u}_{P_0P}$ ($P\in\Omega$) is tangent to the vector $\mathbf{P_0P}$ ($P\in\Omega$) on $\Omega^+_{P_0}$.

(iii) The set $\mathscr{R}=\{r|r=\mathrm{rank}\,\mathscr{A}_n(\mathscr{P}^n)\}$ of all matrices

$$\mathscr{A}_n(\mathscr{P}^n) = \|\Gamma^*_{P_0}(P_0,P_i,P_k)\|, \quad i,k=1,2,\ldots,n \quad n\in\mathbb{N} \tag{5.14}$$

is restricted above

$$\mathrm{rank}\,\mathscr{A}_n(\mathscr{P}^n)\leqslant D, \quad \mathscr{P}^n\subset\Omega^+_{P_0} \tag{5.15}$$

and the supremum D is achieved for the matrix $\mathscr{A}_D(\mathscr{P}^D)$ of the same order D:

$$\det\mathscr{A}_D(\mathscr{P}^D)\neq 0. \tag{5.16}$$

The set $\mathscr{P}^D\subset\Omega^+_{P_0}$ forms a $(D+1)$ point basis in $V^{+*}_{P_0Dt}$.

Let us parametrize the non-null tube $\mathscr{T}_{P_0P}$ by means of a parameter τ in such a way that any point $P(\tau)$ of the section $\mathscr{S}_{1;P(\tau)}(\mathscr{T}_{P_0P})$ corresponds to some value of the parameter τ determined by the relations

$$\left[\sqrt{\sigma(P_0,P(\tau))}+\frac{\tau(1-\tau)}{|\tau(1-\tau)|}\sqrt{\sigma(P(\tau),P)}\right]^2=\sigma(P_0,P),$$
$$\sigma(P_0,P(\tau))=\tau^2\sigma(P_0,P),$$
$$\sigma(P,P(\tau))=(1-\tau)^2\sigma(P_0,P), \tag{5.17}$$
$$\Gamma(P_0,P,(\tau))=2\tau\sigma(P_0,P),$$
$$P(\tau)\in\mathscr{S}_{1;Q(\tau)}(\mathscr{T}_{P_0P}).$$

Here $P(\tau)$ is one of the points of the section $\mathscr{S}_{1;Q(\tau)}(\mathscr{T}_{P_0P})$.

Such a parametrization means that

$$P(\tau)\in\mathscr{T}_{P_0)P}, \quad \tau<0, \quad P_0\in\{P(0)\}, \quad P\in\{P(1)\},$$
$$P(\tau)\in\mathscr{T}_{[P_0P)}, \quad 0\leqslant\tau<1, \quad P(\tau)\in\mathscr{T}_{P_0[P}, \quad 1\leqslant\tau. \tag{5.18}$$

The parametrization is continuous and relations (5.17) for σ satisfy the equation of the tube:

$$F_2(P_0,P,P(\tau)) = \begin{vmatrix} 2\sigma(P_0P) & \Gamma(P_0,P,P(\tau)) \\ \Gamma(P_0,P(\tau),P) & 2\sigma(P_0,P(\tau)) \end{vmatrix} = 0. \tag{5.19}$$

Let us consider the D-dimensional Euclidean space E_D given on the set $\mathbb{R}^D$. The world function has the form (3.4) and the tube ray $\mathscr{T}_{0x}$ can be represented in the form

$$\mathscr{T}_{0x} = \bigcup_{s=0}^{s=D-1}\bigcup_{\tau_s\in\mathbb{R}}\left\{y|y=\tau_0x+\sum_{l=1}^{D-1}a_l(x)\tau_l\right\}, \quad x,y\in\mathbb{R}^D,$$
$$a_l\in\mathbb{R}^D, \quad l=1,2,\ldots,D-1, \quad \tau_s\in\mathbb{R}, \quad s=0,1,\ldots,D-1. \tag{5.20}$$

Here a_l are $D-1$ linear independent vectors which satisfy

$$(a_l(x),a_l(x))=0, \quad (a_l(x),x)=0, \quad l=1,2,\ldots,D-1, \tag{5.21}$$

where

$$(x,y)=\sum_{i,k=1}^{D}g_{ik}x^iy^k, \quad x=\{x^i\}, \quad y=\{y^i\}, \quad l=1,2,\ldots,D.$$

If E_D is the proper Euclidean space, then system (5.21) has only a trivial solution $a_l=0$ ($l=1,2,\ldots,D-1$), and E_D is extremal on $\mathscr{T}_{0x}$. If E_D is the pseudo-Euclidean space of index 1 ($D>2$), then $a_l(x)=0$ ($l=1,2,\ldots,D-1$) for timelike x, $\sigma(0,x)>0$, but $a_l(x)\neq 0$ for spacelike x $(\sigma(0,x)<0)$.

Let us use the curtailing procedure (5.6). Then for the non-null tube one obtains

$$\mathscr{T}^c_{0x} = \{y|\mathscr{T}_{0y} = \mathscr{T}_{0x}\} = \bigcup_{\tau\in\mathbb{R}} \{y|y = \tau x\}, \quad x,y\in\mathbb{R}^D, \tag{5.22}$$

where E_D is extremal on $\mathscr{T}^c_{0x}$.

Thus in the case of the pseudo-Euclidean space the curtailing procedure can be defined as an addition of the curtailing equations $y = \tau x$:

$$\mathscr{T}^c_{0x} = \bigcup_{\tau\in\mathbb{R}} \{y|F_2(0,x,y) = 0 \wedge y = \tau x\}, \quad x,y\in\mathbb{R}^D. \tag{5.23}$$

It is easy to see that the curtailing procedure does not change the extremal timelike tubes, but it does change the nonextremal spacelike tubes.

Theorem 5.1: If the σ space V given on the set Ω determines the D-dimensional σ space $V^{+*}_{P_0Dt}$ of the timelike tangent vector $\mathbf{u}_{P_0P}$ on the set $\Omega^+_{P_0}$ and any timelike direction vector $\mathbf{u}_{P_0P}$ determines the tube ray $\mathscr{T}_{[P_0P}$, then $V^{+*}_{P_0Dt}$ is a D-dimensional Euclidean σ space.

Proof: According to Eq. (5.15) there is a $(D+1)$ point basis $\mathscr{P}^D \subset \Omega^+_{P_0}$ in $V^{+*}_{P_0Dt}$. Let us use the designations

$$g_{ik}(\mathscr{P}^D) = \Gamma^*_{P_0}(P_0,P_i,P_k), \quad i,k = 1,2,...,D, \tag{5.24}$$

$$x_i(P) \equiv (\mathbf{u}_{P_0P_i},\mathbf{P_0P}) = (\mathbf{u}_{P_0P_i},\mathbf{u}_{P_0P}) = \Gamma^*_{P_0}(P_0,P_i,P), \tag{5.25}$$

where $g_{ik}(\mathscr{P}^D)$ is the metric tensor and $x_i(P)$ are covariant coordinates of the point P in the basis $\mathscr{P}^D$ with the basis vectors $\mathbf{e}_i = \mathbf{u}_{P_0P_i}$ $(i = 1,2,...,D)$. According to Eq. (5.13) the vector $\mathbf{u}_{P_0P}$ is supposed to be tangent to $\mathbf{P_0P}$.

Let us write Eq. (5.15) with $n = D+2$, $P_{D+1} = P\in\Omega^+_{P_0}$, $P_{D+2} = Q\in\Omega^+_{P_0}$. Using the designations (5.24) and (5.25) one obtains

$$\operatorname{rank}\begin{matrix} g_{ik}(\mathscr{P}^D) & x_i(P) & x_i(Q) \\ x_k(P) & 2\sigma(P_0,P) & \Gamma^*_{P_0}(P_0,P,Q) \\ x_k(Q) & \Gamma^*_{P_0}(P,Q,P) & 2\sigma(P_0,Q,) \end{matrix} = D, \quad i,k = 1,2,...,D. \tag{5.26}$$

Insofar as the last two columns are linear combinations of the first D columns one obtains

$$\sigma(P_0,P) = \frac{1}{2}\sum_{i=1}^{D} x_i(P)x^i(P), \quad P\in\Omega^+_{P_0}, \tag{5.27}$$

$$\Gamma^*_{P_0}(P_0,P,Q) = \Gamma^*_{P_0}(P_0,Q,P) = \sum_{i=1}^{D} x_i(P)x^i(Q), \quad P,Q\in\Omega^+_{P_0}, \tag{5.28}$$

where

$$x^i(P) = \sum_{k=1}^{D} g^{ik}(\mathscr{P}^D)x_k(P), \quad i = 1,2,...,D. \tag{5.29}$$

Substituting Eqs. (5.27) and (5.28) into Eq. (5.12) one obtains

$$\sigma^*_{P_0}(P,Q) = \frac{1}{2}\sum_{i,k=1}^{D} g_{ik}(\mathscr{P}^D)[x^i(P) - x^i(Q)] \times [x^k(P) - x^k(Q)], \quad P,Q\in\Omega^+_{P_0}. \tag{5.30}$$

The above means that the σ space $V^{+*}_{P_0Dt}$ is Euclidean if there is a one-to-one correspondence between the direction vectors $\mathbf{u}_{P_0P}$ and their coordinates $x(P) = \{x_i(P)\}$, $i = 1,2,...,D$, $x\in\mathbb{R}^D$.

Let $P' = P(\tau)\in\mathscr{T}_{P_0P}$, $Q' = P(\tau')\in\mathscr{T}_{P_0P}$, and $P\in\Omega^+_{P_0}$. Then Eqs. (5.17) and (5.27) lead to

$$\sigma(P',Q') = \frac{1}{2}\sum_{i,k=1}^{D} g^{ik}(\mathscr{P}^D)[x_i(P') - x_i(Q')] \times [x_k(P') - x_k(Q')], \quad P',Q'\in\mathscr{T}_{P_0P}\subset\Omega^+_{P_0}, \tag{5.31}$$

which means that the timelike tubes $\mathscr{T}_{P_0P}$ and $\mathscr{T}^*_{P_0P}$ constructed in the σ spaces V and $V^{+*}_{P_0Dt}$ coincide, respectively, with

$$\mathscr{T}^*_{P_0P} = \mathscr{T}_{P_0P}, \quad \sigma(P,Q) = \sigma^*_{P_0}(P,Q), \quad Q\in\mathscr{T}_{P_0P}. \tag{5.32}$$

Then for $\mathscr{T}_{[P_0P}$ let us take expression (5.23), which is valid for pseudo-Euclidean space,

$$\mathscr{T}^c_{[P_0P} = \bigcup_{\tau>0} \{P'|x(P') = \tau x(P) \wedge 2\tau\sigma(P_0,P) = \Gamma(P_0,P,P')\}, \quad P,P',P_0\in\Omega^+_{P_0}. \tag{5.33}$$

For a timelike tube each of conditions (5.33) is a corollary of the other condition. However, conditions (5.33) will be considered as independent, keeping in mind that we will further use them in the general case.

According to Eq. (2.12) the section of $\mathscr{T}^c_{[P_0P}$ at the point $Q\in\mathscr{T}^c_{[P_0P}$ has the form

$$\mathscr{S}_{1;Q}(\mathscr{T}^c_{[P_0P}) = \bigcup_{\tau>0} \{P'|\sigma(P_0,P') = \sigma(P_0,Q) \wedge \sigma(P,P') = \sigma(P,Q) \wedge x(P') = x(Q) \wedge x(Q) = \tau x(P) \wedge 2\tau\sigma(P_0,P) = \Gamma(P_0,P,Q) \wedge \Gamma(P_0,P,P') = \Gamma(P_0,P,Q)\}. \tag{5.34}$$

As a result of Eqs. (5.19) and (5.27) the first and last of conditions (5.34) are corollaries of the remaining condition and may be omitted. Then

$$\mathscr{S}_{1;Q}(\mathscr{T}^c_{[P_0P}) = \{P'|x(P') = x(Q) \wedge \sigma(P,P') = \sigma(P,Q)\}, \quad Q\in\mathscr{T}^c_{[P_0P}. \tag{5.35}$$

Thus any section corresponds to some value τ. The section is described by

$$x(P') = x(Q) = \tau x(P), \quad x = \{x_i\}, \quad i = 1,2,...,D, \tag{5.36}$$

$$\sigma(P,P') = \sigma(P,Q). \tag{5.37}$$

As a result of the extremality of V on $\mathscr{T}_{[P_0P}$ system

(5.36) and (5.37) has the unique solution $P'=Q$. If Eq. (5.36) has another solution $P'=Q'\neq Q$, then $\sigma(P,Q')\neq\sigma(P,Q)$ and $Q'\notin\mathscr{S}_{1;Q}(\mathscr{T}^c_{[P_0P}$). Besides, Q' does not belong to the other sections, which correspond to other values τ'. Hence, $Q'\notin\mathscr{T}^c_{[P_0P}$. On the other hand, by the supposition of Theorem 5.1 any direction vector $\mathbf{u}_{P_0P}$ determines $\mathscr{T}^c_{[P_0P}$: This means that any solution Q' of Eq. (5.4) belongs to $\mathscr{T}^c_{[P_0P}$. As a result of Eqs. (5.17), (5.27), and the arbitrariness of P, Eq. (5.4) is reduced to Eq. (5.36). Insofar as $Q'\notin\mathscr{T}^c_{[P_0P}$, $P'=Q'$ cannot be a solution of Eq. (5.36). The contradiction obtained shows that there is the unique solution on $P'=Q$ of Eq. (5.36) and a one-to-one correspondence between the point P and its coordinates $x(P)$. Thus Theorem 5.1 has been proved.

In the case of σ space, which is extremal in the zeroth order and does not contain spacelike and null vectors, $\Omega^+_{P_0}=\Omega$ and Theorem 5.1 can be formulated as follows.

Corollary 1: Let the σ space V given on the set Ω be extremal in the zeroth order and $\sigma(P,Q)\geqslant 0$ for any $P,Q\in\Omega$. Let V determine the D-dimensional σ space $V^*_{P_0Dt}$ of the tangent direction vectors $\mathbf{u}_{P_0P}$ on Ω at $P_0\in\Omega$. If any direction vector $\mathbf{u}_{P_0P}$, $P\in\Omega$ determines the tube ray $\mathscr{T}_{[P_0P}$, then $V^*_{P_0Dt}$ is the proper Euclidean σ space.

Using property (5.32), let us try to spread the σ space $V^{+*}_{P_0Dt}$ over the whole set Ω. In the D-dimensional σ space $V^{+*}_{P_0Dt}$ the coordinates (5.25) are defined $\forall P\in\Omega$. Then Eq. (5.25) realizes a mapping $\Omega\to\Omega'_{P_0}\subset\mathbb{R}^D$.

Let us consider the D-dimensional Euclidean σ space $V^{+*}_{P_0Dt}$ given on Ω'_{P_0} by means of the world function (5.30):

$$\sigma^*_{P_0}(x,y)=\frac{1}{2}\sum_{i,k=1}^{D}g^{ik}(\mathscr{P}^D)(x_i-y_i)(x_k-y_k)\,,$$
$$x\in\Omega'_{P_0},\quad y\in\Omega'_{P_0}\,, \tag{5.38}$$

where the dimensionality D is specified by properties of the σ space V mentioned in definitions (5.7)–(5.9). Using expressions (5.33) and (5.35) for the curtailed tube rays in the Euclidean σ space $V^{+*}_{P_0Dt}$ one can use these expressions in the general case. Equations (5.33) and (5.35) are also well defined in the case $\sigma(P_0,P)=0$: They can be considered as the definition of the null curtailed tube $\mathscr{T}^c_{P_0P},\sigma(P_0,P)=0$ and its section at the point $Q\in\mathscr{T}^c_{P_0P}$. In this case the section of $\mathscr{T}^c_{P_0P}$ is defined as a set of points $Q(\tau)\in\mathscr{T}^c_{P_0P}$ with the fixed value τ. Here the null curtailed tube $\mathscr{T}^c_{P_0P}$ is considered as a complex of non-null curtailed tubes $\mathscr{T}^c_{P'_iP}$ with $P'_i\to P_0$ $(i=1,2,\ldots,D)$, $P'_i\in\mathscr{T}_{P_0P_i}$. The curtailed tube ray (5.33) has the definiteness property as a result of its definition by Eq. (5.7). Thus the σ space V is extremal on $\mathscr{T}^c_{[P_0P}$ if

$$\mathscr{S}_{1;Q}(\mathscr{T}^c_{[P_0P})=\{P'|x(P')=x(Q)\wedge\sigma(P,P')=\sigma(P,Q)\}=\{Q\},\quad Q\in\mathscr{T}^c_{[P_0P}\,. \tag{5.39}$$

The curtailed tube rays can be used for the calculation of the limits (5.1) and (5.9), which determine the quantities $(\mathbf{u}_{P_0P},\mathbf{P_0Q})$, $\Gamma^*_{P_0}(P_0,Q,P)$ for the timelike direction vectors $\mathbf{u}_{P_0P}$. Use of the curtailed tube rays enables us to spread definitions (5.4) and (5.6) on the arbitrary direction vectors by replacing $\mathscr{T}_{(P_0Q}$, $\mathscr{T}_{(P_0P}$ by $\mathscr{T}^c_{(P_0Q}$, $\mathscr{T}^c_{(P_0P}$ in Eqs. (5.1) and (5.9).

One should bear in mind that form (5.33) of the curtailed tube supposes a determination of the D-dimensional σ space $V^{+*}_{P_0Dt}$ by the σ space V, although the curtailing operation (5.6) can be used, in principle, in any σ space. If the D-dimensional σ space $V^{+*}_{P_0Dt}$ is determined, then the limits (5.1) and (5.9) take the form

$$(\mathbf{u}_{P_0Q},\mathbf{P_0P})=\lim_{\tau\to+0}\left|\sqrt{\frac{\sigma(P_0,Q)}{\sigma(P_0,Q(\tau))}}\right|\Gamma(P_0,Q(\tau),P)\,,$$
$$x(Q(\tau))=\tau x(Q)\,, \tag{5.40}$$

$$\Gamma^*_{P_0}(P_0,Q,P)=\lim_{\tau\to+0}\lim_{\tau'\to+0}\left|\sqrt{\frac{\sigma(P_0,Q)\sigma(P_0,P)}{\sigma(P_0,Q(\tau))\sigma(P_0,P(\tau'))}}\right|\times\Gamma(P_0,Q(\tau),P(\tau'))\,,\quad x(Q(\tau))=\tau x(Q)\,,$$
$$x(P(\tau'))=\tau' x(P),\quad P,Q,P_0\in\Omega\,. \tag{5.41}$$

For non-null curtailed tubes, when $\mathscr{T}^c_{P_0P}\subseteq\mathscr{T}_{P_0P}$, the limits (5.40) and (5.41) coincide with the limits (5.1) and (5.9). For null curtailed tubes the definition of the limit of Eqs. (5.2) and (5.3) cannot be used. A parametrization of the null curtailed tube ray $\mathscr{T}^c_{P_0P}$ $(\sigma(P_0,P)=0)$, which is used in definitions (5.33) and (5.39), can be obtained as a limit of the parametrization (5.17) of the non-null ray $\mathscr{T}^c_{[P_0P'}$ with $P'\to P, P'\in\mathscr{T}_{[PP'}, \sigma(P,P')\neq 0$. Such a limit is possible in this case only if V is dense on $\mathscr{T}_{P_0}$.

For determination of the σ space $V^*_{P_0Dt}$ on Ω one can use the following procedure, which does not need a density of V at $\forall P\in\mathscr{T}_{P_0}$. Let $\Omega^-_{P_0}=\Omega\setminus\Omega^+_{P_0}$.

(i) $(\mathbf{u}_{P_0Q},\mathbf{P_0P})$, $P\in\Omega$, $Q\in\Omega^+_{P_0}$ is determined by Eqs. (5.1)–(5.3).

(ii) $(\mathbf{u}_{P_0Q},\mathbf{u}_{P_0P})$, $P\in\Omega^+_{P_0}$, $Q\in\Omega^+_{P_0}$ is determined by Eq. (5.9).

(iii) $(\mathbf{u}_{P_0Q},\mathbf{u}_{P_0P})$, $P\in\Omega^-_{P_0}$, $Q\in\Omega^+_{P_0}$ is defined by Eq. (5.13) through $(\mathbf{u}_{P_0Q},\mathbf{P_0P})$, $P\in\Omega^-_{P_0}$, $Q\in\Omega^+_{P_0}$.

(iv) $(\mathbf{u}_{P_0Q},\mathbf{u}_{P_0P})$, $P\in\Omega$, $Q\in\Omega^-_{P_0}$ is determined by Eq. (5.15) through $(\mathbf{u}_{P_0Q},\mathbf{u}_{P_0P})$, $P\in\Omega$, $Q\in\Omega^+_{P_0}$.

(v) $(\mathbf{u}_{P_0Q},\mathbf{P_0P})$, $P\in\Omega$, $Q\in\Omega^-_{P_0}$ is determined by Eq. (5.13) through $(\mathbf{u}_{P_0Q},\mathbf{u}_{P_0P})$, $P\in\Omega$, $Q\in\Omega^-_{P_0}$.

Thus all $(\mathbf{u}_{P_0Q},\mathbf{P_0P})$ and $(\mathbf{u}_{P_0Q},\mathbf{u}_{P_0P})$, $P\in\Omega$, $Q\in\Omega$ are determined through $(\mathbf{u}_{P_0Q},\mathbf{P_0P})$, $P\in\Omega$, $Q\in\Omega^+_{P_0}$ and $(\mathbf{u}_{P_0Q},\mathbf{u}_{P_0P})$, $P\in\Omega^+_{P_0}$, $Q\in\Omega^+_{P_0}$ without using the density of V at $\forall P\in\mathscr{T}_{P_0}$. Practically, it is this procedure that is used for the parametrization (5.33) of the null curtailed tube ray.

Definition 5.10: The σ space V given on the set Ω determines the D-dimensional σ space $V^*_{P_0Dt}$ of the tangent vectors $\mathbf{u}_{P_0P}$ on Ω if the following conditions are fulfilled.

(i) The σ space V is dense at P_0 on any curtailed tube ray $\mathscr{T}^c_{[P_0P}\ni P_0$ and extremal on it.

(ii) Any tube ray $\mathscr{T}^c_{[P_0Q}$ determines the direction vector $\mathbf{u}_{P_0Q}=\mathrm{gen}(\mathscr{T}^c_{[P_0Q})$.

(iii) The scalar product $(\mathbf{u}_{P_0P},\mathbf{u}_{P_0Q})$ is determined and any direction vector $\mathbf{u}_{P_0P}$, $P\in\Omega$ is tangent on Ω to the vector $\mathbf{P_0P}$, $P\in\Omega$ at point P_0.

(iv) The matrix $\mathscr{A}_n(\mathscr{P}^n)$ of any $n+1$ points $\mathscr{P}^n$ satisfies conditions (5.14) and (5.15) with $\mathscr{P}^n \subset \Omega$ instead of $\mathscr{P}^n \subset \Omega^+$.

Theorem 5.2: If the σ space V given on the set Ω determines the D-dimensional σ space $V^*_{P_0 Dt}$ of the tangent vectors $\mathbf{u}_{P_0 P}$ on Ω at P_0 and any direction vector $\mathbf{u}_{P_0 Q}$ determines the curtailed tube ray $\mathscr{T}^c_{[P_0 Q}$ ($\mathscr{T}^c_{[P_0 Q} = \mathrm{gen}(\mathbf{u}_{P_0 Q})$), then the σ space $V^*_{P_0 Dt}$ is Euclidean. The proof of Theorem (5.2) is like that of Theorem 5.1.

Theorem 5.3: Let the σ space V given on the set Ω determine the D-dimensional σ space $V^*_{P_0 Dt}$ of the tangent direction vectors $\mathbf{u}_{P_0 P}$ on Ω and let any direction vector $\mathbf{u}_{P_0 Q}$ determine the curtailed tube ray $\mathscr{T}^c_{[P_0 Q}$. If V is dense at any point $P \in \Omega$, then V determines the D-dimensional manifold on the open set $\Omega_0 = \bar{\Omega} \setminus B$, where $\bar{\Omega}$ is the closure of Ω and B is the boundary of $\bar{\Omega}$.

Proof: According to Theorem 5.2 there is a one-to-one mapping $\Omega \to \mathscr{M} \subset \mathbb{R}^D$, where $\mathbb{R}^D$ is the space of all coordinates $x = \{x_i(P)\}$, $(i = 1,2,...,D)$. Let us define the δ vicinity of the point $x \in \mathscr{M}$ as a set of $y \in \mathscr{M}$ satisfying the condition

$$|x-y|^2 = \sum_{i=1}^{D} (x_i - y_i)^2 < \delta, \quad \delta > 0 .$$

Then the set $\mathscr{M}$ is dense at any point (i.e., any δ vicinity of $x \in \mathscr{M}$ contains at least one point $y \neq x$) because the σ space V is dense at any point P on any curtailed tube ray $\mathscr{T}^c_{P_0 P}$. Let $\bar{\mathscr{M}}$ be the closure of $\mathscr{M}$. Here $\bar{\mathscr{M}}$ does not contain isolated points because $\mathscr{M}$ is dense at any point. Let us remove all boundary points of $\bar{\mathscr{M}}$. Then $\bar{\mathscr{M}}$ transforms into an open region $\mathscr{M}_0$ of $\mathbb{R}^D$: $\mathscr{M} \subset \mathscr{M}_0 \subset \mathbb{R}^D$.

Let us use the one-to-one correspondence between Ω and $\mathscr{M}$ and construct the mappings $\bar{\Omega} \leftrightarrow \bar{\mathscr{M}}$ and $\Omega_0 \leftrightarrow \mathscr{M}_0$ with $\Omega \subset \Omega_0$. The σ function on the Ω_0 can be defined from the world function on Ω by means of the proper limiting process. Then the σ space V_0 on Ω_0 arises. The set Ω_0 with the coordinate system defined by Eq. (5.25) is a manifold. Theorem (5.3) has been proved.

VI. THE RIEMANNIAN SPACE

Definition 6.1: The D-dimensional Riemannian space V is a D-dimensional manifold $\mathscr{M}$ with the quadratic form

$$(dS)^2 = g_{ik}(x)\,dx^i\,dx^k \tag{6.1}$$

given at any point P of $\mathscr{M}$ in some coordinate system K on $\mathscr{M}$. Here $x = \{xi\}$, $(i = 1,2,...,D)$ are contravariant coordinates of point P in the coordinate system K; g_{ik}, $(i,k = 1,2,...,D)$ is the metric tensor. Here and below the summation is made over repeated superscripts and subscripts from 1–D.

In the thus-defined Riemannian space V the world function $\sigma(P,P')$ is defined through the interval

$$S(P,P') = S(x,x') = \int_{P'}^{P} \sqrt{g_{ik}\,dx^i\,dx^k} \tag{6.2}$$

by means of the relation

$$\sigma(x,x') = \sigma(P,P') = \tfrac{1}{2} S^2(P,P') . \tag{6.3}$$

Integration in Eq. (6.2) is produced along the geodesic $\mathscr{L}_{P'P}$, which is an extremal of the integral (6.2) considered as a functional of $x^i = x^i(\tau)$, $(i = 1,2,...,D)$. The functions $x^i(\tau)$ satisfy

$$\frac{d^2x^i}{d\tau^2} + \gamma^i_{kl}(x)\frac{dx^k}{d\tau}\frac{dx^l}{d\tau} = 0, \quad i = 1,2,...,D , \tag{6.4}$$

$$\gamma^i_{kl}(x) = \frac{1}{2} g^{is}\left(\frac{\partial g_{sk}}{\partial x^l} + \frac{\partial g_{sl}}{\partial x^k} - \frac{\partial g_{kl}}{\partial x^s}\right), \quad i,k,l = 1,2,...,D , \tag{6.5}$$

$$g = \det\|g_{ik}\|, \quad g_{ik}\frac{dx^i}{d\tau}\frac{dx^k}{d\tau} = \mathrm{const} . \tag{6.6}$$

The rather small region Ω of the Riemannian space V is considered, so that one and only one geodesic passes through two different points P, $P' \in \Omega$, $P \neq P'$.

If the metric (6.1) is definite, i.e., if

$$g_{ik}x^i x^k = 0$$

has the unique solution $x^i = 0 (i = 1,2,...,D)$, then the Riemannian σ space V with the σ function (6.2) and (6.3) is extremal in the zeroth and first orders. For the opposite case the σ space V is nonextremal in the zeroth order, but it can be extremal in the first order (on the curtailed tubes).

We shall consider only the cases when the Riemannian space is locally either proper Euclidean or pseudo-Euclidean of index 1. In the latter case the spacelike geodesics are curtailed geodesic tubes of the form (5.33). In both cases the σ space V is thought to be extremal in the first order.

The σ function defined by Eqs. (6.2) and (6.3) satisfies[4]

$$\sigma_i g^{ik}(x)\sigma_k = 2, \quad \sigma(x,x') = \sigma(x',x), \sigma(x,x) = 0 , \tag{6.7}$$

where $g^{ik}(x)$ is the metric tensor at point P with the coordinates x and

$$\sigma_i \equiv \frac{\partial \sigma}{\partial x^i}, \quad \sigma_{i'} \equiv \frac{\partial \sigma}{\partial x^{i'}}, \quad i,i' = 1,2,...,D . \tag{6.8}$$

The prime at the index shows that the differentiation is produced with respect to the coordinates x' of point P'. The absence of the prime shows that the differentiation is produced with respect to the coordinates x of point P. Essentially, Eq. (6.7) is a corollary of the fact that integration in Eq. (6.2) is produced along the extremal.

There is another formulation of the extremal property (6.7) which does not contain the metric tensor explicitly: It is valid for any Riemannian space. This formulation has the form of a system of differential equations containing only the σ function and its derivatives.[5,6]

$$\sigma_i \sigma^{i,k'} \sigma_{k'} = 2\sigma, \quad \sigma(x,x') = \sigma(x',x), \quad \sigma(x,x) = 0, \tag{6.9}$$

$$G_{ik\|l} = 0, \quad i,k,l = 1,2,...,D, \tag{6.10}$$

where $\sigma^{i,k'}$ is determined by

$$\sigma^{i,k'}\sigma_{l,k'} = \delta^i_l, \quad i,l = 1,2,...,D, \tag{6.11}$$

$$\sigma_{l,k'} \equiv \frac{\partial^2 \sigma(x,x')}{\partial x^l\,\partial x'^{k'}}, \quad l,k' = 1,2,...,D. \tag{6.12}$$

The symbol $(\,)_{\|l}$ denotes the tangent derivative with respect to x^l, i.e., the covariant derivative with respect to x^l with the Christoffel symbol

$$\Gamma^i_{kl}(x,x') = \sigma^{ij'}\sigma_{klj'}, \quad \sigma_{klj'} \equiv \frac{\partial^3\sigma}{\partial x^k\,\partial x^l\,\partial x'^{j'}}, \quad i,k,l,j' = 1,2,...,D, \tag{6.13}$$

and G_{ik} is defined by the relation

$$G_{ik} = G_{ik}(x,x') \equiv \sigma_{i\|k} = \frac{\partial\sigma_i}{\partial x^k} - \Gamma^l_{ik}\,\sigma_l, \quad i,k = 1,2,...,D. \tag{6.14}$$

Equations (6.9) and (6.10) are corollaries of Eq. (6.7), but Eq. (6.7) can be obtained as a corollary of Eqs. (6.9) and (6.10) and the "boundary" condition[5,6]

$$[\sigma_{i,k'}] = \sigma_{i,k'}(x,x')|_{x'=x} = -g_{ik}(x), \quad i,k = 1,2,...,D. \tag{6.15}$$

The tensor G_{ik} is the metric tensor at point x of the Euclidean space $E_{x'}$, which is tangent to the Riemannian space V at point x'. The geodesic mapping $V \to E_{x'}$, is produced in such a way that any geodesic $\mathscr{L}_{P'P}$ of V passing through point P' is mapped into a straight line of $E_{x'}$ tangent to $\mathscr{L}_{P'P}$ at point P'. At such a mapping the length of any intercept of the geodesic $\mathscr{L}_{P'P}$ and the angles between the geodesics at point P' are conserved. The coordinate system K in V is mapped into the coordinate system $K_{x'}$ in $E_{x'}$. Here G_{ik} is the metric tensor of $E_{x'}$ in the coordinate system $K_{x'}$. In $E_{x'}$ the σ function of the arguments P and P'' has the form[5,6]

$$\sigma^*_{P'}(P,P'') = \sigma^*_{x'}(x,x'') = \sigma(x',x) + \sigma(x',x'') - \sigma_{i'}(x',x)g^{i'k'}(x')\sigma_{k'}(x',x''). \tag{6.16}$$

Relation (6.16) sets in correspondence the world function $\sigma^*_{P'}$ of the Euclidean space $E_{x'}$ to a point P' and the world function σ of the Riemannian space V.

Theorem 6.1: Let the σ space V given on the D-dimensional manifold $\mathscr{M}$ have the world function σ, which is the twice-differentiable function of coordinates. Then the σ space determines a D-dimensional Riemannian space R on $\mathscr{M}$.

Proof: Expanding $\sigma(x,x+dx)$ into a series over the powers of dx^i, one obtains, as a result of properties (2.1) of the σ function,

$$\tfrac{1}{2}(dS)^2 = \sigma(x,x+dx) = \tfrac{1}{2}g_{ik}(x)dx^i\,dx^k + o(|dx|^2),$$

$$g_{ik}(x) = \left[\frac{\partial\sigma(x,y)}{\partial y^i\,\partial y^k}\right]_{y=x}, \quad i,k = 1,2,...,D. \tag{6.17}$$

Theorem (6.1) has been proved.

Remark: In general, the σ function σ_R of the Riemannian space R, defined by Eqs. (6.2)–(6.5) and (6.17), does not coincide with the world function σ of the σ space V.

Definition 6.2: The σ space V given on a set Ω_0 is a Riemannian σ space if Ω_0 is a subset of points of a Riemannian space R and the σ function σ_R of R, defined by Eqs. (6.2)–(6.5) and (6.17), coincides with the world function σ of V:

$$\sigma(P_0,P) = \sigma_R(P_0,P), \quad P_0 \in \Omega_0, \quad P \in \Omega_0. \tag{6.18}$$

Theorem 6.2: Let the σ space V be given on the set Ω and let the following conditions be fulfilled.

(i) Here V determines the D-dimensional σ space $V^*_{P_0,Dt}$ of the tangent direction vectors $\mathbf{u}_{P_0P}$ at any point $P_0 \in \Omega$ on Ω.

(ii) Each direction vector $\mathbf{u}_{P_0P}$ determines the curtailed tube ray $\mathscr{T}^c_{[P_0P}$. If V is dense at any point $P \in \Omega$, then the σ space V_0 given on $\Omega_0 = \bar{\Omega}\backslash B$ is the Riemannian space and the world function defined by Eqs. (6.2)–(6.5) and (6.17) coincides with the σ function of V_0.

Proof: According to Theorem 5.3 the σ space V determines the D-dimensional manifold on the set $\Omega_0 = \bar{\Omega}\backslash B$, where B is the boundary of $\bar{\Omega}$, and the σ space V_0 on Ω_0 arises. Let P_0, $P \in \Omega_0$ be arbitrary points of Ω_0. According to Eqs. (5.27) and (5.29) the σ function is the twice-differentiable function of the coordinates (5.25). As a result of Theorem 6.1 the σ space V_0 determines a D-dimensional Riemannian space R on Ω_0, with the metric tensor (5.24) at $P_0 \in \Omega_0$ in the coordinate system K_{P_0} defined by Eq. (5.25). Four σ spaces arise on Ω_0: V_0, $V^*_{P_0,Dt}$, R, and $E_{x'}$, with the corresponding σ functions σ, $\sigma^*_{P_0}$, σ_R, $\sigma^*_{RP_0}$. According to Eqs. (5.32) and (6.16) one obtains

$$\sigma(P_0,P) = \sigma^*_{P_0}(P_0,P), \quad \sigma_R(P_0,P) = \sigma^*_{RP_0}(P_0,P), \quad P_0,P \in \Omega_0, \quad x'_i = x_i(P_0), \quad x_i = x_i(P), \quad i = 1,2,...,D. \tag{6.19}$$

According to Eq. (6.17) two Euclidean σ spaces $V^*_{P_0,Dt}$ and $E_{x'}$ coincide in the infinitesimal vicinity of point P_0; hence, they coincide everywhere:

$$\sigma^*_{P_0}(P_0,P) = \sigma^*_{RP_0}(P_0,P), \quad P_0 \in \Omega_0, \quad P \in \Omega_0. \tag{6.20}$$

Equations (6.20) and (6.19) lead to Eq. (6.18). The σ functions of V_0 and R coincide. The σ space V_0 given on Ω_0 is a Riemannian space. Curtailed tubes in V_0 coincide with the geodesics in R $\mathscr{T}^c_{P_0P} = \mathscr{L}_{P_0P}$.

Theorem 6.3: The σ space V given on a D-dimensional manifold $\mathscr{M}$ is a Riemannian σ space if V is extremal on any curtailed tube $\mathscr{T}^c_{P_0P}$, P_0, $P \in \mathscr{M}$ and the σ function is the twice-differentiable function of coordinates.

Proof: One can verify that all suppositions of Theorem 6.2 are fulfilled as a result of the suppositions of Theorem 6.3. Then Theorem 6.3 is valid as a result of Theorem 6.2.

VII. VIOLATION OF EXTREMALITY IN THE FIRST ORDER

A σ space defined on a manifold and extremal in all orders $n \geqslant 1$ is a Euclidean σ space. According to Theorem 6.3 a σ space defined on a manifold and extremal in the first order is a Riemannian space. Which are properties of a σ space defined on a manifold, but nonextremal in the first order? Can such a σ space have a bearing on real event space?

The real event space is usually considered as a four-dimensional pseudo-Euclidean space of index 1 or as a four-dimensional Riemannian space. In both cases the event space considered as σ space is extremal in the first order (on curtailed tubes) and timelike tubes coincide with timelike curtailed tubes. The world line of a free particle placed at point x' and having the four-velocity u^i is described by an algebraic equation with respect to $x(\tau)$:

$$\sigma_{i'}(x,x') = u_{i'}\tau = g_{i'k'}(x')u^{k'}\tau, \quad i' = 0,1,2,3, \tag{7.1}$$

where τ is a parameter along the world line. If $\det\|\sigma_{i,k}\| \neq 0$, then Eq. (7.1) can be solved with respect to x^i: It describes a one-dimensional world line

$$x^i = x^i(\tau), \quad i = 0,1,2,3. \tag{7.2}$$

Equation (7.1) (there are four) is distinguished from

$$F_2(P_0,P_1,P) = 0 \tag{7.3}$$

in the respect that Eq. (7.1) always describes the one-dimensional line; however, Eq. (7.3) does that only in the case when the extremality conditions are fulfilled.

The circumstance that the σ function is both the transformation function describing the classical particle motion and the world function describing the event space properties permits us to use test particle observations for the determination and description of event space properties. In a certain sense the motion of the free classical particles and event space properties are identical because each is described through the other.

The drawing of a geodesic is a way of describing the space properties, but it is equivalent to the observation of free classical particle motion. Pointlike particles are necessary for a test of space properties at small distances. However, pointlike particles are simultaneously the particles of a small mass (electrons, protons, etc.) which move according to quantum mechanics laws.

If one describes the quantal motion of microparticles in terms of Feynman path integrals, then the particle moves along arbitrary trajectories, but not only along extremal ones, although the motion along them is most probable. This means a violation of extremality in the sense of property (i) of Sec. II and does not permit us to use quantal particles for testing the space-time properties, as one can do by means of classical particles.

Another approach is possible using the following hypothesis: *The space-time considered as σ space is not extremal in the first order, i.e., a tube of lines (not geodesic) passes through any two points of the space-time. The microparticle motion is described by this tube* (i.e., by the world tube and not by a world line).

The above hypothesis permits us to use microparticle motion for testing of the space-time properties and, in particular, for determination of the extremality violation. The section ($t = \text{const}$) of a world tube is in general a surface (string), but not a point. This circumstance is associated with the string model of elementary particles, which is currently popular.

Example 6: Let there be a coordinate system K on the four-dimensional manifold. Let the σ function have in this coordinate system K the form

$$\sigma(P,P') = \sigma(x,x') = \tfrac{1}{2} q(1 + (\varepsilon q/l^2)^2,$$
$$q = (x - x')^2 = (t - t')^2 - (\mathbf{x} - \mathbf{x}')^2, \tag{7.4}$$

where $x = (t,\mathbf{x})$, $x' = (t',\mathbf{x}')$ are coordinates of the points P and P', respectively. Here l is some characteristic length and ε is a small parameter $|\varepsilon| \ll 1$ describing a small violation of extremality.

Let $P_0 = (0,\mathbf{0})$, $P_1 = (a,\mathbf{0})$, and $\varepsilon a^2/l^2 \ll 1$. The equation for the timelike tube $\mathscr{T}_{P_0P_1}$ has the approximate form

$$\mathbf{x}^2 = -6\frac{\varepsilon t^2(t-a)^2}{l^2} + O(\varepsilon^2), \quad \varepsilon\frac{t^2}{l^2} \ll 1, \quad \varepsilon\frac{(t-a)^2}{l^2} \ll 1. \tag{7.5}$$

If $\varepsilon = 0$, then the tube is a geodesic $\mathbf{x} = 0$. If $\varepsilon > 0$, then the timelike tube $\mathscr{T}_{P_0P_1}$ degenerates into two points $\mathscr{T}_{P_0P_1} = \{P_0,P_1\}$, but formally the σ space remains extremal on the tube $\mathscr{T}_{P_0P_1}$ because extremal properties degenerate into the trivial form. If $\varepsilon < 0$, then the timelike straight line $\mathscr{L}_{P_0P_1}$ transforms into a three-dimensional surface. The extremal properties of definiteness and minimal section are violated. The tube is more close to the geodesic the less $\sqrt{\varepsilon}t/l$.

If, indeed, real space-time is distinguished from Riemannian space at a small distance, then one should expect that attempts to describe the particle motion in terms of world lines leads to contradictions and difficulties. In the nonrelativistic approximation these difficulties have been successfully handled in terms of a probabilistic description of quantum mechanics; however, one cannot be sure that it is the best way of overcoming these difficulties.

[1]H. S. Ruse, Proc. London Math. Soc. **32**, 87 (1931).
[2]H. S. Ruse, Q. J. Math. **2**, 190 (1931).
[3]J. L. Synge, Proc. London Math. Soc. **32**, 241 (1931).
[4]J. L. Synge, *Relativity: The General Theory* (North-Holland, Amsterdam, 1960).
[5]Yu. A. Rylov, Izv. Vyssh. Ucheb. Zaved., Mat. No. 3, 132 (1962), in Russian.
[6]Yu. A. Rylov, Ann. Phys. (Leipzig) **12**, 329 (1964).
[7]G. E. Gorelik, in *Einsteinovskiyi Sbornik 1984–1985*, edited by X. Kobsarev (Nauka, Moscow, 1988), in Russian.
[8]S. M. Christensen, Phys. Rev. D **14**, 2490 (1976).
[9]A. O. Barvinsky and G. A. Vilkovisky, Phys. Rep. **119**, 1 (1985).

Quantum mechanics as an infinite-dimensional Hamiltonian system with uncertainty structure: Part I

Renzo Cirelli, Alessandro Manià, and Livio Pizzocchero
Dipartimento di Fisica, Università di Milano, Milano, Italy and Istituto Nazionale di Fisica Nucleare, Sezione di Milano, Milano, Italy

(Received 22 August 1989; accepted for publication 27 June 1990)

Schrödinger quantum mechanics is formulated as an infinite-dimensional Hamiltonian system whose phase space carries an additional structure (uncertainty structure) to account for the probabilistic character of the theory. The algebra of observables is described as an algebra of smooth functions on the quantal phase space, with a product naturally induced by the geometrical structures living on that manifold. The possibility of generalizing Schrödinger mechanics along these lines is discussed.

I. INTRODUCTION

Pure states in ordinary (Schrödinger) quantum mechanics can be given a fairly natural geometrical description as the points of an infinite-dimensional Kähler manifold, the projective space $\mathbf{P}(\mathscr{H})$ of the Hilbert space $\mathscr{H}$ of the system, as a matter of fact.[1-5]

This geometrical description is interesting in many respects. An ordinary quantum system appears as an infinite-dimensional Hamiltonian system of a special kind, in which the Kähler metric plays a central role. In this paper, we want to make clear this role: Indeed, we show that, apart from providing the symplectic structure through its fundamental form, the metric gives the "dispersion structure" of the system, therefore taking charge of the probabilistic character of quantum mechanics. (The term "uncertainty structure" appearing in the title refers precisely to this structure; it is closer to physical intuition, but less precise from a technical viewpoint.)

We further exploit the geometrical description of Schrödinger mechanics looking at the observables. The observables are represented in this framework by a special, quite natural class of real-valued smooth functions on $\mathbf{P}(\mathscr{H})$ that we call *Kählerian functions*. Obviously enough, we look for a suitable algebraic structure for Kählerian functions and we find that it is intimately connected with Kähler properties and, in particular, with the constancy of the holomorphic sectional curvature of the state manifold of the system. By the way, this throws further light on the relation between the holomorphic sectional curvature and Planck's constant.

On these grounds, rather naturally, we are led to the question of possible generalizations of Schrödinger quantum mechanics obtained assuming as state manifold a significant substitute for $\mathbf{P}(\mathscr{H})$. The answer is that, under some general conditions which appear reasonable from the physical viewpoint, one is necessarily led to Schrödinger mechanics.

Our paper is organized as follows. Section II provides the basic mathematical tools, defining a number of bilinear composition laws between the smooth functions on a general almost Kählerian manifold and introducing Kählerian functions. Section III contains the Kähler formulation of Schrödinger mechanics on $\mathbf{P}(\mathscr{H})$. Section IV discusses the possibility to generalize Schrödinger mechanics.

II. PRELIMINARIES: $*_\nu$ PRODUCT AND KÄHLERIAN FUNCTIONS

In order to standardize the language, we start recalling a few definitions. Given a real, smooth (i.e., C^∞) Banach manifold $\mathscr{M}$, we shall use notations such as $\mathscr{L}(T\mathscr{M})$, $\mathscr{L}(T^*\mathscr{M}, T\mathscr{M})$, $\mathscr{L}_n(T\mathscr{M}, \mathbf{R})$, etc., to denote the vector bundles of base $\mathscr{M}$ whose fibers at any point $\mathbf{x}$ of $\mathscr{M}$ are $\mathscr{L}(T_{\mathbf{x}}\mathscr{M})$ (the Banach space of bounded linear operators on the tangent space at $\mathbf{x}$), $\mathscr{L}(T^*_{\mathbf{x}}\mathscr{M}, T_{\mathbf{x}}\mathscr{M})$ (the Banach space of bounded linear maps from the cotangent to the tangent space at $\mathbf{x}$), $\mathscr{L}_n(T_{\mathbf{x}}\mathscr{M}, \mathbf{R})$ (the Banach space of bounded n-linear forms on $T_{\mathbf{x}}\mathscr{M}$), and so on.

An almost complex structure on $\mathscr{M}$ is a smooth section J of $\mathscr{L}(T\mathscr{M})$ such that $J^2 = -1$. Such a J is called integrable if its torsion[6] is zero; i.e., if for every open set $\mathscr{U}$ of $\mathscr{M}$ and every pair X, Y of smooth vector fields on $\mathscr{U}$,

$$[JX,JY] - [X,Y] - J[X,JY] - J[JX,Y] = 0. \quad (2.1)$$

An almost Kählerian manifold is a triple $(\mathscr{M},J,g)$, where $\mathscr{M}$ is a real, smooth Hilbertian manifold, J is an almost complex structure, and g is a Kähler metric, i.e., a Riemann metric[7] such that:

(1) g is J invariant:

$$g_{\mathbf{x}}(J_{\mathbf{x}}X_{\mathbf{x}}, J_{\mathbf{x}}Y_{\mathbf{x}}) = g_{\mathbf{x}}(X_{\mathbf{x}}, Y_{\mathbf{x}}), \quad \forall \mathbf{x} \in \mathscr{M}, X_{\mathbf{x}}, Y_{\mathbf{x}} \in T_{\mathbf{x}}\mathscr{M}; \quad (2.2)$$

(2) the fundamental form of the metric, that is the nondegenerate two-form ω defined by

$$\omega_{\mathbf{x}}(X_{\mathbf{x}}, Y_{\mathbf{x}}) := g_{\mathbf{x}}(J_{\mathbf{x}}X_{\mathbf{x}}, Y_{\mathbf{x}}), \quad \forall \mathbf{x} \in \mathscr{M}, X_{\mathbf{x}}, Y_{\mathbf{x}} \in T_{\mathbf{x}}\mathscr{M} \quad (2.3)$$

is closed.

Note that an almost Kählerian manifold is canonically a symplectic manifold. If J is integrable, we shall say that $(\mathscr{M},J,g)$ is a Kähler manifold.

Now let us fix an almost Kählerian manifold $(\mathscr{M},J,g)$. The fundamental form ω and the Kähler metric g induce for every $\mathbf{x} \in \mathscr{M}$ two top-linear isomorphisms $I_{\mathbf{x}}$, $G_{\mathbf{x}}$ between the cotangent space $T^*_{\mathbf{x}}\mathscr{M}$ and the tangent space $T_{\mathbf{x}}\mathscr{M}$. These are defined implicitly by

$$\omega_{\mathbf{x}}(I_{\mathbf{x}}\alpha_{\mathbf{x}}, X_{\mathbf{x}}) = \langle \alpha_{\mathbf{x}}, X_{\mathbf{x}} \rangle, \quad g_{\mathbf{x}}(G_{\mathbf{x}}\alpha_{\mathbf{x}}, X_{\mathbf{x}}) = \langle \alpha_{\mathbf{x}}, X_{\mathbf{x}} \rangle, \quad (2.4)$$

 0022-2488/90/122891-07$03.00

for $\alpha_x \in T_x^* \mathscr{M}$, $X_x \in T_x \mathscr{M}$. We shall denote by I, G the smooth sections $x \mapsto I_x$, $x \mapsto G_x$ of $\mathscr{L}(T^*\mathscr{M}, T\mathscr{M})$; one easily checks that $G = J \circ I$.

Definition 2.1: Let $f,l \in C^\infty(\mathscr{M},\mathbf{R})$. The Poisson and Riemann brackets of f and l, denoted by $\{f,l\}$ and $((f,l))$, respectively, are the real smooth functions on $\mathscr{M}$ defined by

$$\{f,l\} := \langle df, I\, dl \rangle \tag{2.5}$$

and

$$((f,l)) := \langle df, G\, dl \rangle. \tag{2.6}$$

By bilinearity, the definition of Poisson and Riemann brackets will be extended to smooth, complex valued functions on $\mathscr{M}$.

The definition of Poisson bracket given in (2.5) is standard. The second bracket operation is less usual; we call it Riemann bracket, because its definition depends only on g. Note that Eq. (2.4) and the J invariance of ω and g allow us to reformulate equivalently the definitions (2.5) and (2.6) as follows:

$$\{f,l\} = \omega(I\, df, I\, dl) = \omega(G\, df, G\, dl); \tag{2.5'}$$

$$((f,l)) = g(G\, df, G\, dl) = g(I\, df, I\, dl). \tag{2.6'}$$

We now define other bilinear composition laws between smooth functions on $\mathscr{M}$. The physical relevance of these composition laws in connection with quantum mechanics will appear in the next section.

Definition 2.2: Let $f,l \in C^\infty(\mathscr{M},\mathbf{C})$. The Kähler bracket of f and l, denoted by $<f,l>$, is the smooth complex valued function on $\mathscr{M}$ defined by

$$<f,l> := ((f,l)) + i\{f,l\}. \tag{2.7}$$

For $\nu \in \mathbf{R}$, the $\circ_\nu$ product and the $*_\nu$ product of f and l are the smooth complex valued functions on $\mathscr{M}$ defined, respectively, by

$$f \circ_\nu l := \tfrac{1}{2}\nu((f,l)) + fl, \tag{2.8}$$

$$f *_\nu l := \tfrac{1}{2}\nu <f,l> + fl. \tag{2.9}$$

When equipped with Riemann bracket, or, alternatively, with $\circ_\nu$ product, $C^\infty(\mathscr{M},\mathbf{R})$ [resp. $C^\infty(\mathscr{M},\mathbf{C})$] becomes a real (resp. complex) commutative and nonassociative algebra. The Kähler bracket and the $*_\nu$ product provide two more structures of complex, nonassociative algebra for $C^\infty(\mathscr{M},\mathbf{C})$. With any of these two bilinear composition laws and with the natural involution defined by complex conjugation, $C^\infty(\mathscr{M},\mathbf{C})$ is an involutive algebra.

We also observe that, for $f,l \in C^\infty(\mathscr{M},\mathbf{C})$,

$$f *_\nu l = f \circ_\nu l + (i/2)\nu\{f,l\}, \tag{2.10}$$

$$f \circ_\nu l = \tfrac{1}{2}(f *_\nu l + l *_\nu f), \tag{2.11}$$

$$\{f,l\} = (1/i\nu)(f *_\nu l - l *_\nu f). \tag{2.12}$$

Definition 2.3: Let $f \in C^\infty(\mathscr{M},\mathbf{R})$, and let X be the vector field $I\, df$. We shall say that the function f is Kählerian if

$$L_X g = 0,$$

where L_X denotes the Lie derivative along X. More generally, if $f \in C^\infty(\mathscr{M},\mathbf{C})$, we shall say that f is Kählerian if $\mathrm{Re}\, f$ and $\mathrm{Im}\, f$ are Kählerian. The set of Kählerian functions will be denoted by $\mathscr{K}(\mathscr{M},\mathbf{R})$ or $\mathscr{K}(\mathscr{M},\mathbf{C})$.

We think it is convenient to add a few comments about the definition of real Kählerian functions. First of all, in the language of symplectic manifolds, $X = I\, df$ is the Hamiltonian vector field corresponding to f. The condition $L_X g = 0$ tells us that the integral flow of X, that is the Hamiltonian flow of f, preserves the metric g. Note that, by the Liouville theorem, we have automatically $L_X \omega = 0$. From this also follows $L_X J = 0$ [remember that J is uniquely determined by ω and g via Eq. (2.3)]. Therefore, if f is Kählerian, the Hamiltonian flow of f preserves the whole Kähler structure. We remark that if the distance induced by g on every connected component of $\mathscr{M}$ is complete the Hamiltonian flow of any real Kählerian function is globally defined on $\mathbf{R} \times \mathscr{M}$; in fact, it is known that infinitesimal isometries of complete Riemann manifolds are complete.[8]

It is easy to check that $\mathscr{K}(\mathscr{M},\mathbf{R})$ [resp. $\mathscr{K}(\mathscr{M},\mathbf{C})$] is a Lie subalgebra of $C^\infty(\mathscr{M},\mathbf{R})$ [resp. $C^\infty(\mathscr{M},\mathbf{C})$] with respect to Poisson bracket. The behavior of Kählerian functions with respect to the other bilinear composition laws introduced in this section will be studied extensively in the rest of the paper.

III. THE KÄHLER FORMULATION OF SCHRÖDINGER QUANTUM MECHANICS

Let us now specialize our considerations to a particular Kähler manifold, namely the projective space $\mathbf{P}(\mathscr{H})$, where $\mathscr{H}$ is a complex Hilbert space with scalar product $(\,|\,)$.

Here, $\mathbf{P}(\mathscr{H})$ is defined as the set of one-dimensional subspaces, or rays, of $\mathscr{H}$; for every $x \in \mathscr{H} \setminus \{0\}$, $[x]$ will denote the ray through x. If $\mathscr{H}$ is the Hilbert space of a Schrödinger quantum system, the rays of $\mathscr{H}$ represent the pure states of the system and $\mathbf{P}(\mathscr{H})$, once equipped with its natural differential structure (see below), can be regarded as the "state manifold" of the quantum system, just like phase space is the state manifold of a classical mechanical system.

The structure carried by the quantal state manifold is richer than the structure of classical phase space. The latter is a symplectic manifold; this is true also of $\mathbf{P}(\mathscr{H})$, but in this case the symplectic form is the fundamental form of a Kähler metric. The almost complex structure and the Kähler metric carried by $\mathbf{P}(\mathscr{H})$ are not natural objects for classical phase space, or, at least, they are not essential in the Hamiltonian formulation of classical mechanics. On the contrary, these objects are fundamental for the formulation of quantum mechanics illustrated in this paper.

Let us illustrate the Kähler structure of $\mathbf{P}(\mathscr{H})$. Following Ref. 1, we first define on $\mathbf{P}(\mathscr{H})$ a structure of holomorphic manifold by means of the atlas $\{(\mathscr{V}_h, b_h, \mathscr{C}_h)\}(h \in \mathscr{H}, \|h\| = 1)$, where, for every h, $(\mathscr{V}_h, b_h, \mathscr{C}_h)$ is the chart with domain $\mathscr{V}_h$, and local model the complex Hilbert space $\mathscr{C}_h$, defined as follows:

$$\mathscr{V}_h := \{[x] \in \mathscr{H} \mid (h\,|\,x) \neq 0\}; \tag{3.1}$$

$$\mathscr{C}_h := [h]^\perp; \tag{3.2}$$

$$b_h : \mathscr{V}_h \to \mathscr{C}_h, \quad [x] \mapsto b_h([x]) := x/(h\,|\,x) - h \tag{3.3}$$

[in Eq. (3.2), $^\perp$ denotes the orthogonal complement].

A holomorphic manifold can be regarded in an obvious way as a real smooth manifold with an integrable almost

complex structure. This will be, from now on, our attitude towards $\mathbf{P}(\mathscr{H})$. As a real manifold, $\mathbf{P}(\mathscr{H})$ is coordinatized by the atlas $\{(\mathscr{V}_h, \mathscr{R}\circ b_h, \mathscr{R}\mathscr{E}_h)\}$, $(h\in\mathscr{H}, \|h\|=1)$, where $\mathscr{R}\mathscr{E}_h$ denotes the realification of $\mathscr{E}_h$ (the real Hilbert space obtained from $\mathscr{E}_h$ restricting to $\mathbf{R}$ the field of scalars) and $\mathscr{R}:\mathscr{E}_h\to\mathscr{R}\mathscr{E}_h$, $v\mapsto\mathscr{R}_v$ is the canonical bijection. Vectors of $\mathscr{R}\mathscr{E}_h$ will be always represented in the form $\mathscr{R}v$, where $v\in\mathscr{E}_h$.

We now write down the local expression with respect to a chart $(\mathscr{V}_h, \mathscr{R}\circ b_h, \mathscr{R}\mathscr{E}_h)$ of the Kähler metric g carried by $\mathbf{P}(\mathscr{H})$. The metric, g is a smooth section of $\mathscr{L}_2(T\mathbf{P}(\mathscr{H}), \mathbf{R})$; its local expression is the smooth map

$$g^h:\mathscr{R}\mathscr{E}_h\to\mathscr{L}_2(\mathscr{R}\mathscr{E}_h,\mathbf{R}),\quad \mathscr{R}z\mapsto g^h_{\mathscr{R}z}$$

where

$$g^h_{\mathscr{R}z}(\mathscr{R}v,\mathscr{R}w) := 2\nu\,\mathrm{Re}\left(\frac{(v|w)}{1+\|z\|^2}-\frac{(v|z)(z|w)}{(1+\|z\|^2)^2}\right). \tag{3.4}$$

In Eq. (3.4), ν denotes an arbitrarily chosen positive constant, whose role will be clear shortly afterwards. For a characterization of the distance induced by g, see Ref. 9.

As is usual, we shall denote by ω the fundamental form of the metric. If we regard ω as a section of $\mathscr{L}_2(T\mathbf{P}(\mathscr{H}), \mathbf{R})$, its local expression in any chart $(\mathscr{V}_h, \mathscr{R}\circ b_h, \mathscr{R}\mathscr{E}_h)$ is the map

$$\omega^h:\mathscr{R}\mathscr{E}_h\to\mathscr{L}_2(\mathscr{R}\mathscr{E}_h,\mathbf{R}),\quad \mathscr{R}z\mapsto\omega^h_{\mathscr{R}z},$$

where

$$\omega^h_{\mathscr{R}z}(\mathscr{R}v,\mathscr{R}w) := 2\nu\,\mathrm{Im}\left(\frac{(v|w)}{1+\|z\|^2}-\frac{(v|z)(z|w)}{(1+\|z\|^2)^2}\right). \tag{3.5}$$

The Kähler metric (3.4) on $\mathbf{P}(\mathscr{H})$ is the infinite-dimensional generalization of the well-known Fubini–Study metric[10] on $\mathbf{P}(\mathbf{C}^n)$; we shall therefore call it in the same way. Its holomorphic sectional curvature is constant and equal to $2/\nu$.

It must be remarked that projective spaces are, up to isomorphisms, the only connected, simply connected and complete Kähler manifolds of constant and positive holomorphic sectional curvature; this follows from the Hawley–Igusa theorem[6] and will be fundamental in the sequel.

We also give the local expressions of the sections I, G of $\mathscr{L}(T^*\mathbf{P}(\mathscr{H}), T\mathbf{P}(\mathscr{H}))$ defined by Eq. (2.4). For every chart $(\mathscr{V}_h, \mathscr{R}\circ b_h, \mathscr{R}\mathscr{E}_h)$, and for $\mathscr{R}z\in\mathscr{R}\mathscr{E}_h$, $\alpha\in(\mathscr{R}\mathscr{E}_h)^*$,

$$I^h_{\mathscr{R}z}(\alpha) = (1/2\nu)(1+\|z\|^2)\mathscr{R}(-iu(\alpha)-i(z|u(\alpha))z), \tag{3.6}$$

$$G^h_{\mathscr{R}z}(\alpha) = (1/2\nu)(1+\|z\|^2)\mathscr{R}(u(\alpha)+(z|u(\alpha))z), \tag{3.7}$$

where $u(\alpha)$ is the unique element of $\mathscr{E}_h$ such that

$$\langle\alpha,\mathscr{R}v\rangle = \mathrm{Re}(u(\alpha)|v),\quad \forall v\in\mathscr{E}_h. \tag{3.8}$$

Up to now, we have studied the differential structure of the quantum state space $\mathbf{P}(\mathscr{H})$ and the geometrical structure it carries without any reference to dynamics. In Schrödinger quantum mechanics, the dynamics of the system is determined by the map

$$\mathbf{R}\times\mathbf{P}(\mathscr{H})\to\mathbf{P}(\mathscr{H}),\quad (t,[x])\mapsto[e^{-(i/\hbar)tH}x],$$

where H is a (typically unbounded) self-adjoint operator in $\mathscr{H}$. To implement this fact into our geometrical language, we shall fix our attention on Kähler isomorphisms of $\mathbf{P}(\mathscr{H})$ onto itself [i.e., smooth diffeomorphisms $\Phi:\mathbf{P}(\mathscr{H})\to\mathbf{P}(\mathscr{H})$ with the properties $\Phi^*J=J$ and $\Phi^*g=g$].

If U is any unitary operator on $\mathscr{H}$, the map $\mathbf{P}(\mathscr{H})\to\mathbf{P}(\mathscr{H})$, $[x]\mapsto[Ux]$ is a Kähler isomorphism of $\mathbf{P}(\mathscr{H})$. Conversely,[1] any Kähler isomorphism of $\mathbf{P}(\mathscr{H})$ is induced by a unitary operator U, unique up to a phase factor.

Furthermore,[2] for every self-adjoint operator A in $\mathscr{H}$, possibly unbounded, the family of maps $(\Phi_t)_{(t\in\mathbf{R})}$, where

$$\Phi_t:\mathbf{P}(\mathscr{H})\to\mathbf{P}(\mathscr{H}),\quad [x]\mapsto[e^{-itA}x]$$

is a continuous one parameter group of Kähler isomorphisms of $\mathbf{P}(\mathscr{H})$ and, vice versa, every continuous one parameter group $(\Phi_t)_{(t\in\mathbf{R})}$ of Kähler isomorphisms of $\mathbf{P}(\mathscr{H})$ is induced by a self-adjoint operator A [boundedness of A amounts to smoothness for $(\Phi_t)_{(t\in\mathbf{R})}$].

Therefore, in the present framework, the dynamical law for a quantum system is described as a continuous one parameter group of Kähler isomorphisms of the state manifold $\mathbf{P}(\mathscr{H})$. Note that Kähler isomorphisms are also isomorphisms for the symplectic structure defined by the fundamental form of the metric. So, in this approach a quantum system is (also) a Hamiltonian system, in the same way as classical conservative mechanical systems, if, as a Hamiltonian system, a pair is meant consisting of a symplectic manifold and a one parameter group of symplectic isomorphisms. In particular, a quantum system has a Hamiltonian function generating the dynamics; on this we shall return while discussing observables.

Observables and their algebraic structure are the third constituent of Schrödinger mechanics. In the traditional language, they are described by self-adjoint operators; ideally, we can suppose that every self-adjoint operator represents an observable. To simplify our treatment, we shall consider only *bounded* self-adjoint operators. We shall see that they are in one-to-one correspondence with the real Kählerian functions.

Definition 3.1: Let A be a bounded linear operator on $\mathscr{H}$. We shall denote by $\langle A\rangle$ the mean value function of A, defined by

$$\langle A\rangle:\mathbf{P}(\mathscr{H})\to\mathbf{C},\quad [x]\mapsto\langle A\rangle_{[x]} := (x|Ax)/\|x\|^2, \tag{3.9}$$

and by $\Delta^2 A$ the squared dispersion function of A, defined by

$$\Delta^2 A:\mathbf{P}(\mathscr{H})\to\mathbf{C},\quad [x]\mapsto\Delta^2_{[x]}A := \langle(A-\langle A\rangle_{[x]})^2\rangle_{[x]}. \tag{3.10}$$

The maps (3.9) and (3.10) are smooth. If A is self-adjoint, $\langle A\rangle$ is real valued, $\Delta^2 A$ is nonnegative, and we can define the dispersion function $\Delta A := \sqrt{\Delta^2 A}$; the physical interpretation of $\langle A\rangle$ and ΔA in quantum mechanics is well known.

Again supposing A self-adjoint, we give the local expressions $\langle A\rangle^h:\mathscr{E}_h\to\mathbf{R}$ and $(d\langle A\rangle)^h:\mathscr{E}_h\to(\mathscr{E}_h)^*$ of $\langle A\rangle$ and $d\langle A\rangle$ with respect to a chart $(\mathscr{V}_h, \mathscr{R}\circ b_h, \mathscr{R}\mathscr{E}_h)$; for $\mathscr{R}z$, $\mathscr{R}v\in\mathscr{R}\mathscr{E}_h$ one has

$$\langle A\rangle^h(\mathscr{R}z) = (z+h|A(z+h))/(1+\|z\|^2); \tag{3.11}$$

$$\langle (d\langle A\rangle)^h_{\mathscr{R}z}, \mathscr{R}v\rangle = 2\,\mathrm{Re}\left(\frac{1}{1+\|z\|^2}A(z+h) - \frac{(h\,|A(z+h))}{1+\|z\|^2}h - \frac{(A(z+h)|z+h)}{(1+\|z\|^2)^2}z\,\middle|\,v\right). \tag{3.12}$$

Furthermore, the local expressions $X^h: \mathscr{R}\mathscr{C}_h \to \mathscr{R}\mathscr{C}_h$, $Y^h: \mathscr{R}\mathscr{C}_h \to \mathscr{R}\mathscr{C}_h$ of the vector fields $X := Id\langle A\rangle$ and $Y := Gd\langle A\rangle$ are

$$X^h(\mathscr{R}z) = (1/\nu)\mathscr{R}(i(h\,|A(z+h))(z+h) - iA(z+h)) \tag{3.13}$$

and

$$Y^h(\mathscr{R}z) = (1/\nu)\mathscr{R}(-(h\,|A(z+h))(z+h) + A(z+h)). \tag{3.14}$$

Using local expressions, we can prove the following.

Proposition 3.2: Let A be a bounded self-adjoint operator on $\mathscr{H}$. Then the flow of the vector field $X = Id\langle A\rangle$ is complete and is given by the one parameter group $(\Phi_t)_{(t\in\mathbf{R})}$, where

$$\Phi_t([x]) = [e^{-i(t/\nu)A}x], \quad \forall [x]\in\mathbf{P}(\mathscr{H}). \tag{3.15}$$

The previous proposition first appeared in Ref. 11. See also Refs. 1, 2, 4, and 12.

Proposition 3.3: Let f be a complex-valued function on $\mathbf{P}(\mathscr{H})$. Then f is Kählerian if and only if there is a bounded operator A such that $f = \langle A\rangle$.

Proof: This is just a reformulation of a result obtained in Ref. 2. ■

Let us discuss briefly the meaning of Proposition 3.3, restricting our considerations to real Kählerian functions. According to 3.3, the observables of a quantum system, represented by self-adjoint operators in the traditional language, can be described in this approach by a *selected* class of real smooth functions on the state manifold, namely the Kählerian functions. Since these are exactly the functions whose Hamiltonian flow preserves the geometric structures carried by $\mathbf{P}(\mathscr{H})$, we can say that observables are the generators of the structural symmetries of the state manifold.

It must also be noticed that Proposition 3.2 throws further light on the dynamical law for the system. Let us consider for a moment the idealized case of a system whose Hamiltonian operator H is bounded (the unbounded case can be treated using a more sophisticated technique, see Ref. 2). Proposition 3.2 tells us that the general solution of Schrödinger's equation induced by H is the Hamiltonian flow of the function $(\nu/\hbar)\langle H\rangle$. If we want the Hamiltonian function for the system to be just $\langle H\rangle$, we must stipulate that

$$\nu = \hbar. \tag{3.16}$$

This equation provides a geometrical interpretation of Planck's constant in terms of the curvature of the quantal state manifold; such an interpretation has been pointed out in Refs. 3 and 4.

We now examine the connection between the algebraic structure of $\mathscr{L}(\mathscr{H})$ and the different bilinear composition laws between smooth functions on $\mathbf{P}(\mathscr{H})$ defined as in the previous section. The introduction of these composition laws, and the discussion of their links with operator algebras and with the geometry of the manifold on which they work are the main results of the paper.

Proposition 3.4: Let $A, B\in\mathscr{L}(\mathscr{H})$. Then, with ν as in Eq. (3.4),

(a) $\{\langle A\rangle,\langle B\rangle\} = \langle (1/i\nu)[A,B]\rangle$;
(b) $((\langle A\rangle,\langle B\rangle)) = (1/\nu)\langle AB+BA\rangle - (2/\nu)\langle A\rangle\langle B\rangle$
and, in particular
$((\langle A\rangle,\langle A\rangle)) = (2/\nu)\Delta^2 A$;
(c) $\langle\langle A\rangle,\langle B\rangle\rangle = (2/\nu)(\langle AB\rangle - \langle A\rangle\langle B\rangle)$;
(d) $\langle A\rangle\circ_\nu\langle B\rangle = \frac{1}{2}\langle AB+BA\rangle$;
(e) $\langle A\rangle *_\nu \langle B\rangle = \langle AB\rangle$.

Proof: Once we have proved (a) and (b), the remaining statements follow trivially. Furthermore, by linearity arguments, we can limit ourself to prove (a) and (b) in the case A and B are self-adjoint. The proof can be obtained going to local expressions. By definition, $\{\langle A\rangle,\langle B\rangle\} = \langle d\langle A\rangle, X\rangle$ and $((\langle A\rangle,\langle B\rangle)) = \langle d\langle A\rangle, Y\rangle$, where $X := Id\langle B\rangle$ and $Y := Gd\langle B\rangle$.

Using Eqs. (3.12)–(3.14), we obtain for the local expressions of the Poisson and Riemann brackets

$$\{\langle A\rangle,\langle B\rangle\}^h(\mathscr{R}z) = \frac{(z+h\,|(1/i\nu)(AB-BA)(z+h))}{1+\|z\|^2};$$

$$((\langle A\rangle,\langle B\rangle))^h(\mathscr{R}z) = \frac{1}{\nu}\frac{(z+h\,|(AB+BA)(z+h))}{1+\|z\|^2} - \frac{2}{\nu}\frac{(z+h\,|A(z+h))}{1+\|z\|^2}\times\frac{(z+h\,|B(z+h))}{1+\|z\|^2}.$$ ■

We remark that one of the statements contained in Proposition 3.4, namely (a), was already known (see, for example, Ref. 11). Proposition 3.4 can be used for a Kähler formulation of Schrödinger quantum mechanics; it leads us to a number of conclusions that are reported hereafter.

(1) If we put $\nu = \hbar$ in (a), we find a precise mathematical support for the statement that "$(1/i\hbar)[\,,\,]$ is the quantum analog of Poisson bracket."

(2) The Riemann bracket is the operation we must use in this framework to compute the dispersion of observables. Putting $\nu = \hbar$ in (b), we obtain that, for every observable $f\in\mathscr{K}(\mathbf{P}(\mathscr{H}),\mathbf{R})$ and every state $[x]\in\mathbf{P}(\mathscr{H})$, the results of a great number of measurements of f in the state $[x]$ are distributed with standard deviation $\sqrt{\frac{1}{2}\hbar((f,f))([x])}$ around the mean value $f([x])$. This also explains the role of the Riemannian structure carried by the quantum state manifold: *It is the structure that takes charge of the probabilistic character of quantum mechanics.*

(3) Point (d) tells us that $\circ_\nu$ -product between (real or complex) Kählerian functions is the equivalent in this framework of Jordan product between operators. In particular, $\circ_\nu$ product provides the power structure for observables, since, for every $A\in\mathscr{L}(\mathscr{H})$,

$$\langle A^n\rangle = (\cdots((\langle A\rangle\circ_\nu\langle A\rangle)\circ_\nu\langle A\rangle)\cdots\circ_\nu\langle A\rangle) \quad (n \text{ times}).$$

(4) Point (e) tells us that $*_\nu$ product between complex-valued Kählerian functions is the equivalent of the operator product. Apart from applications to quantum mechanics, this result is by itself interesting because it allows us to formulate a functional representation theorem for the algebra $\mathscr{L}(\mathscr{H})$. In order to refine the functional representation, let us put

$$\|f\|_\nu := \sqrt{\sup_{[x]\in\mathbf{P}(\mathscr{H})} (\bar{f}*_\nu f)([x])}. \tag{3.17}$$

The map $f \mapsto \|f\|_\nu$ is a norm on the linear space of the smooth functions f such that the right side of Eq. (3.17) is finite. Equipped with $*_\nu$ product, with complex conjugation and with the norm $\|\ \|_\nu$, $\mathscr{K}(\mathbf{P}(\mathscr{H}),\mathbf{C})$ is a W^* algebra, and the map $A \mapsto \langle A \rangle$ is an isomorphism between the W^* algebras $\mathscr{K}(\mathbf{P}(\mathscr{H}),\mathbf{C})$ and $\mathscr{L}(\mathscr{H})$. Starting from this fact, it is possible to develop a general functional representation theory for C^* algebras,[13] generalizing the classical spectral representation for commutative C^* algebras. In this general theory, the Kähler manifold $\mathbf{P}(\mathscr{H})$ is replaced by a topological fiber bundle in which every fiber is a Kähler manifold isomorphic to a projective space. It is known[9] that such a structure of "Kähler bundle" is naturally admitted by the set of pure states of any C^* algebra.

Once discovered the link between the operator formalism on $\mathscr{H}$ and the "Kählerian functions formalism" on $\mathbf{P}(\mathscr{H})$, to translate a notion from one language to the other is just an exercise; in particular, this operation can be performed for the notion of spectrum of a bounded self-adjoint operator A on $\mathscr{H}$. Indeed, it is not difficult to prove[14] that a nonzero vector x in $\mathscr{H}$ is an eigenvector of A if and only if $d_{[x]}\langle A \rangle = 0$ or, equivalently, if and only if $[x]$ is a fixed point for the vector field $Id\ \langle A \rangle$; in this case, the corresponding eigenvalue is $\langle A \rangle_{[x]}$. By a suitable generalization, it is possible to treat along the same lines the continuous spectrum.

We conclude the section with an observation about the limiting behavior as $\nu \to 0^+$ of the formalism we have developed; by Eq. (3.16), this can be interpreted as the classical limit of quantum mechanics. Now, for $\nu \to 0^+$ the holomorphic sectional curvature of the Fubini–Study metric grows to $+\infty$, or, in a more pictorial language, the "radius of holomorphic sectional curvature" tends to zero. Thus we can say that in our approach the classical limit is a singular limit in which the quantum state manifold, in a sense, collapses.

IV. GEOMETRICAL QUANTUM MECHANICS?

The results obtained in the previous section show that one can formulate Schrödinger quantum mechanics using the language of differential geometry instead of the traditional apparatus based on Hilbert spaces and linear operators. Since projective spaces are, up to isomorphisms, the only connected, simply connected and complete Kähler manifolds of constant and positive holomorphic sectional curvature, we can give the following geometrical definition of Schrödinger quantum system:

Definition 4.1: A Schrödinger quantum system is a quadruple $(\mathscr{M},J,g,(\Phi_t)_{(t\in\mathbf{R})})$, where $(\mathscr{M},J,g)$ is a connected, simply connected and complete Kähler manifold of constant holomorphic sectional curvature $2/\hbar$, and $(\Phi_t)_{(t\in\mathbf{R})}$ is a continuous one parameter group of Kähler isomorphisms of $(\mathscr{M},J,g)$.

Given such a structure, we interpret $\mathscr{M}$ as the set of pure states of a physical system, the real Kählerian functions as the observables, etc., and $(\Phi_t)_{(t\in\mathbf{R})}$ as the dynamical law. We remark that such a geometrical attitude toward quantum mechanics agrees with the general scheme of quantum theory proposed in Ref. 15 (pure states as the points of a manifold and observables as a *selected* class of real functions on the pure states, in contrast to classical mechanics where *any* real function on the phase space is an observable).

One may wonder whether it is possible, on the grounds of some reasonable assumption, to enlarge Definition 4.1 so as to include some hypothetical "generalized quantum system." Why should the state manifold be isomorphic to a projective space? Is Kähler structure really essential for a satisfactory definition of quantum system, or can it be replaced by some more general geometrical structure? Of course, to discuss these problems, we must necessarily put some *a priori* limitation on the most general structure we are willing to consider. Keeping in mind what we have learned in the previous section working $\mathbf{P}(\mathscr{H})$, we make the following choice.

Definition 4.2: A trial quantum system (TQS) is a quadruple $(\mathscr{M},\omega,g,(\Phi_t)_{(t\in\mathbf{R})})$ where $\mathscr{M}$ is a real, smooth Hilbertian manifold, and ω, g are, respectively, a symplectic form and a Riemann metric on $\mathscr{M}$; $(\Phi_t)_{(t\in\mathbf{R})}$ is a continuous one-parameter group of isomorphisms of $(\mathscr{M},\omega,g)$:

$$\Phi_t^*\omega = \omega, \quad \Phi_t^* g = g.$$

Given a TQS, we shall denote by $\mathscr{K}(\mathscr{M},\mathbf{R})$ the set of real smooth functions f on $\mathscr{M}$ such that the Hamiltonian vector field $X = I\,df$ preserves g: $L_X g = 0$. $\mathscr{K}(\mathscr{M},\mathbf{C})$ will be the set of complex-valued smooth functions f on $\mathscr{M}$ such that $\mathrm{Re}\, f,\ \mathrm{Im}\, f \in \mathscr{K}(\mathscr{M},\mathbf{R})$.

In the previous definition, the term "trial" emphasizes the fact that we do not know whether every structure of that kind can be reasonably interpreted as the geometrical description of a quantum system. Apart from this, the ideas underlying Definition 4.2 are clear:

(1) $\mathscr{M}$ is the set of pure states of the system: We require it admits a symplectic structure because we want a Hamiltonian theory.

(2) $\mathscr{K}(\mathscr{M},\mathbf{R})$ is the set of observable functions, characterized as in the case of Schrödinger quantum systems as the generators of the symmetries of the geometrical structures carried by $\mathscr{M}$. For every observable f and every state $\mathbf{x}\in\mathscr{M}$, the mean value obtained by a great number of measurements of f in the state $\mathbf{x}$ is $f(\mathbf{x})$.

(3) The Riemann metric g gives the "dispersion structure" of the system: For $f\in\mathscr{K}(\mathscr{M},\mathbf{R})$ and $\mathbf{x}\in\mathbf{M}$, the dispersion of f in the state $\mathbf{x}$ is

$$\Delta_{\mathbf{x}} f := \sqrt{\tfrac{1}{2}\hbar((f,f))(x)}, \tag{4.1}$$

where $((\ ,\ ))$ denotes the Riemann bracket operation, defined as in Eq. (2.6).

Like Eq. (2.6), Eqs. (2.5) and (2.7)–(2.9) also have a well-defined meaning when applied to a TQS; we shall use them to extend to general trial quantum systems the definitions of all bilinear composition laws introduced in Sec. II. The symbols $\circ_\hbar$ and $*_\hbar$ will be employed to denote $\circ_\nu$ product and $*_\nu$ product with $\nu = \hbar$. For future application, it will also be useful the following.

Definition 4.3: Given a TQS, we shall say that $\mathscr{K}(\mathscr{M},\mathbf{R})$ is *full* if for every $x\in M$ the differentials $d_x f$, $f\in\mathscr{K}(\mathscr{M},\mathbf{R})$, span the whole cotangent space $T_x^*\mathscr{M}$.

The fullness condition means that the set of observables is sufficiently rich; to ask it is reasonable from a physical viewpoint. This condition is satisfied by Schrödinger quantum systems, as well as by other trial quantum systems, for example those in which $\mathscr{M}$ is the Grassmannian of d-planes in a complex Hilbert space ($d = 2,3,\ldots$), g is the natural Kähler metric of the Grassmannian and ω is its fundamental form.

Having delimited by Definition 4.2 the general geometrical structure allowed in our discussion, we pass to examine the questions we formulated at the beginning of this section. A possible answer to such questions is the one provided in Refs. 3 and 4, that, in our setting, sounds as follows: Schrödinger quantum systems are the only trial quantum systems in which the state manifold satisfies a number of geometrical requirements (in particular, homogeneity, isotropy, and positivity of the curvature; simple connectedness is also explicitly asked). In the approach we shall follow here, attention is fixed on the fullness condition, the measurement process and the algebraic structure of observables. In our opinion, this approach is better founded from a physical viewpoint: Rather than putting *by hand* a set of geometrical conditions on the state manifold, it selects indirectly the admissible geometries through a careful discussion of the observables and their behavior. In this discussion of the observables, fullness is the assumption in which, at a first sight, the connection with the geometry of the state manifold is most direct;[16] however, the fullness requirement alone is not resolutive, and the conclusions of our analysis will depend strongly on the other assumptions we will make.

We now start discussing the question: Is Kähler structure really necessary to build quantum mechanics? Given any TQS, we have a unique smooth section J of $\mathscr{L}(T\mathscr{M})$ such that

$$\omega_x(X_x,Y_x) = g_x(J_xX_x,Y_x), \quad \forall x\in\mathscr{M}, X_x,Y_x\in T_x\mathscr{M}. \tag{4.2}$$

From the nondegeneracy of ω follows that J_x is a top-linear isomorphism for every $x\in\mathscr{M}$, If $J^2 = -1$, J is an almost complex structure and the triple $(\mathscr{M},J,g)$ is an almost Kählerian manifold.

As we shall now see, there is a relevant physical reason to ask that $J^2 = -1$: This condition is equivalent to a sufficiently strong formulation of the uncertainty principle. The need for a strong formulation of the uncertainty principle was first pointed out in a different context by Ref. 17. In the present, geometrical framework, we are led to the following.

Definition 4.4: We say that a TQS $(\mathscr{M},\omega,g,(\Phi_t)_{(t\in\mathbf{R})})$ satisfies the uncertainty principle if the following conditions are satisfied for every $x\in M$:

(a) for every $f,l\in\mathscr{K}(\mathscr{M},\mathbf{R})$,

$$\Delta_x f\Delta_x l \geq \tfrac{1}{2}\hbar|\{f,l\}(x)|;$$

(b) for every $f\in\mathscr{K}(\mathscr{M},\mathbf{R})$,

$$\Delta_x f = \mathrm{Inf}\{\lambda\in[0,+\infty]) \mid \lambda\Delta_x l \geq \tfrac{1}{2}\hbar|\{f,l\}(x)| \quad \forall l\in\mathscr{K}(\mathscr{M},\mathbf{R})\}.$$

Proposition 4.5: Let $(\mathscr{M},\omega,g,(\Phi_t)_{(t\in\mathbf{R})})$ be a TQS in which $\mathscr{K}(\mathscr{M},\mathbf{R})$ is full. Then $(\mathscr{M},\omega,g,(\Phi_t)_{(t\in\mathbf{R})})$ satisfies the uncertainty principle if and only if the tensor field J defined by Eq. (4.2) is an almost complex structure.

Proof: Let us fix $x\in M$. For $f,l\in\mathscr{K}(\mathscr{M},\mathbf{R})$, we have

$$\Delta_x f = \sqrt{\tfrac{1}{2}\hbar g_x(J_xX_x,J_xX_x)}, \quad \Delta_x l = \sqrt{\tfrac{1}{2}\hbar g_x(J_xY_x,J_xY_x)}$$
$$\{f,l\}(x) = \omega_x(X_x,Y_x),$$

where we have put $X_x := I_x d_x f$, $Y_x := I_x d_x l$. From here and the fullness hypothesis follows that conditions (a) and (b) are equivalent to

(a′) $g_x(J_xX_x,J_xX_x)g_x(J_xY_x,J_xY_x) \geq \omega_x^2(X_x,Y_x)$ for every X_x, $Y_x\in T_x\mathscr{M}$;

(b′) for every $X_x\in T_x\mathscr{M}$

$$g_x(J_xX_x,J_xX_x) = \mathrm{Inf}\{\Lambda\in[0,+\infty) \mid \Lambda g_x(J_xY_x,J_xY_x) \geq \omega_x^2(X_x,Y_x)\ \forall Y_x\in T_x\mathscr{M}\}.$$

Moreover, $J_x^2 = -1$ if and only if

(c) $g_x(J_xX_x,J_xX_x) = g_x(X_x,X_x)$ for every $X_x\in T_x\mathscr{M}$.

Thus we must show that conditions (a′) and (b′) together are equivalent to (c).

Assume (a′) and (b′) are satisfied. Then, for every $X_x\in T_x\mathscr{M}$

$$g_x^2(X_x,X_x) = g_x^2(J_xJ_x^{-1}X_x,X_x) = \omega_x^2(J_x^{-1}X_x,X_x) \leq g_x(X_x,X_x)g_x(J_xX_x,J_xX_x),$$

so that

$$g_x(X_x,X_x) \leq g_x(J_xX_x,J_xX_x). \tag{4.3}$$

On the other hand, since for every $\xi\in\mathbf{R}$, X_x, $Y_x\in T_x\mathscr{M}$, one has

$$0 \leq g_x(X_x+\xi J_xY_x,X_x+\xi J_xY_x) = g_x(X_x,X_x) + \xi^2 g_x(J_xY_x,J_xY_x) - 2\xi\omega_x(X_x,Y_x),$$

choosing $\xi = \omega_x(X_x,Y_x)/g_x(J_xY_x,J_xY_x)$, we obtain

$$g_x(X_x,X_x)g_x(J_xY_x,J_xY_x) - \omega_x^2(X_x,Y_x) \geq 0.$$

Therefore, from (b′), we have

$$g_x(X_x,X_x) \geq g_x(J_xX_x,J_xX_x). \tag{4.4}$$

Conversely, let (c) be satisfied. Then using Schwarz's inequality, we easily infer (a′). To obtain (b′), assume we are given, for a fixed $X_x\in T_x\mathscr{M}$, a non-negative real number Λ such that

$$\Lambda g_x(J_xY_x,J_xY_x) \geq \omega_x^2(X_x,Y_x), \quad \forall Y_x\in T_x\mathscr{M}. \tag{4.5}$$

In particular, this inequality will hold for $Y_x = J_xX_x$. In this case, using (c) we get

$$\Lambda g_x(J_xX_x,J_xX_x) \geq g_x^2(J_xX_x,J_xX_x)$$

that is $\Lambda \geq g_x(J_xX_x,J_xX_x)$. This proves (b′). ■

Proposition 4.5 gives us sufficient motivations for re-

stricting our considerations to trial quantum systems of *almost Kählerian type*, i.e., those in which the tensor field J defined by Eq. (4.2) is an almost complex structure.

We now examine the behavior of observable functions with respect to the bilinear composition laws induced by g, ω. Let us take a function $f \in \mathscr{K}(\mathscr{M}, \mathbf{R})$; then, the equation

$$(f \circ_\hbar f)(\mathbf{x}) = \Delta_x^2 f + f^2(\mathbf{x}), \tag{4.6}$$

which follows immediately from Eq. (4.1) and the definition of $\circ_\hbar$ product, supports the interpretation of $f \circ_\hbar f$ as the observable "square of f" (remember of the identity $\langle F^2 \rangle = \Delta^2 F + \langle F \rangle^2$, holding in probability theory for any square-summable random variable F). Therefore, it is natural to ask that

$$f \in \mathscr{K}(\mathscr{M}, \mathbf{R}) \Rightarrow f \circ_\hbar f \in \mathscr{K}(\mathscr{M}, \mathbf{R}).$$

This amounts to ask that

$$f, l \in \mathscr{K}(\mathscr{M}, \mathbf{C}) \Rightarrow f *_\hbar l \in \mathscr{K}(\mathscr{M}, \mathbf{C})$$

[remember that $*_\hbar = \circ_\hbar + (i/2)\hbar\{\ ,\ \}$ and that the Kählerian functions are closed with respect to Poisson bracket].

The $*_\hbar$ product couples together the "power structure" for observables, respented by $\circ_\hbar$ product, with the Lie algebra structure, defined by the Poisson bracket, in a way which is suitable to build a complex, associative algebra. Now, the following proposition holds.

Proposition 4.6: Let $(\mathscr{M}, \omega, g, (\Phi_t)_{(t \in \mathbf{R})})$ be a TQS of almost Kählerian type in which $\mathscr{K}(\mathscr{M}, \mathbf{R})$ is full. Let us consider the pair of conditions:

(a) $f *_\hbar l \in \mathscr{K}(\mathscr{M}, \mathbf{C})$ for every $f, l \in \mathscr{K}(\mathscr{M}, \mathbf{C})$;

(b) $(f *_\hbar l) *_\hbar k = f *_\hbar (l *_\hbar k)$ for every $f, l, k \in \mathscr{K}(\mathscr{M}, \mathbf{C})$.

These conditions are satisfied if and only if the almost complex structure J defined by Eq. (4.2) is integrable and $(\mathscr{M}, J, g)$ is a Kähler manifold of constant holomorphic sectional curvature $2/\hbar$.

Proof: It will result from the theorems we will establish in an accompanying paper,[18] devoted to the study of properties of Kählerian functions in general, almost Kählerian manifolds. We shall see that condition (b) is equivalent to integrability for the almost complex structure and that, taken for granted integrability, the apparently "innocuous" condition (a) is satisfied if and only if the curvature is constant and equals $2/\hbar$. ■

Propositions 4.5 and 4.6 provide a possible answer to the questions formulated in this section. Summing up, we can say that if a trial quantum system has the following properties then the space state is a Kähler manifold of constant holomorphic sectional curvature $2/\hbar$: (i) the observables are full; (ii) the uncertainty principle is satisfied; and (iii) the observables form an associative algebra with respect to $*_\hbar$ -product. Under the additional assumptions of connectedness, simple connectedness and completeness, this necessarily leads us to ordinary Schrödinger mechanics.

Note added in proof: After this work was submitted for publication, a paper of S. Weinberg appeared ["Testing Quantum Mechanics," Ann. Phys. **194**, 336 (1989)], concerning the general foundations of quantum mechanics and its possible nonlinear generalizations. Here, a $*$-product is introduced on the quantum observables, in a framework that is slightly different from ours, clashing with the lack of associativity in the nonlinear case. We think the general geometrical setting proposed in our paper is relevant in connection with these problems.

[1] R. Cirelli, P. Lanzavecchia, and A. Manià, "Normal pure states of the Von Neumann algebra of bounded operators as Kähler manifold," J. Phys. A **16**, 3829 (1983).

[2] R. Cirelli and P. Lanzavecchia, "Hamiltonian vector fields in quantum mechanics," Nuovo Cimento B **79**, 271 (1984).

[3] A. Heslot, "Une caractérization des espaces projectifs complexes," C. R. Acad. Sci. Paris **298**, 95 (1984).

[4] A. Heslot, "Quantum mechanics as a classical theory," Phys. Rev. D **31**, 1341 (1985).

[5] R. Cirelli and L. Pizzocchero, "On the integrability of quantum mechanics as an infinite-dimensional Hamiltonian system," to appear in Nonlinearity.

[6] S. Kobayashi and K. Nomizu, *Foundations of Differential Geometry* (Wiley, New York, 1969), Vols. I and II.

[7] Throughout the paper, by Riemann metric, we shall mean a *strong* Riemann metric. Analogously, by symplectic form, we shall always mean a *strong* symplectic form.

[8] The quickest proof is obtained as follows: If X is a smooth vector field such that $L_X g = 0$, for every integral curve λ of X having as domain a bounded open interval (a,b), the norm $\sqrt{g(X,X)}$ is constant along λ, and thus $d(\lambda(t), \lambda(t')) \leqslant \text{const} \times |t - t'|$ for $t, t' \in (a,b)$. From here and the completeness of the distance d induced by g, it follows that $\lambda(t)$ admits well defined limits $\lambda(a)$, $\lambda(b)$ as t tends to a, b, respectively. Now, using the existence theorem for differential equations, we can extend λ in a left neighborhood $(a - \epsilon, a]$ of a and in a right neighborhood $[b, b + \eta)$ of b. In conclusion, every integral curve of X with domain a bounded open interval can be extended. On the completeness of infinitesimal isometries, see also Ref. 6.

[9] M. C. Abbati, R. Cirelli, P. Lanzavecchia, and A. Manià, "Pure states of general quantum-mechanical systems as Kähler bundles," Nuovo Cimento B **83**, 43 (1984).

[10] G. Fubini, "Sulle metriche definite da una forma Hermitiana," Atti Ist. Veneto **6**, 501 (1903); E. Study, "Kürzeste Wege im komplexen Gebiete," Math. Ann. **60**, 321 (1905).

[11] R. Hermann, "Topics in the mathematics of quantum mechanics," in *Interdisciplinary Mathematics* (Mathematical Science, Brookline, MA, 1973), Vol. VI. Here and in Ref. 12, attention is mainly fixed on $\mathbf{P}(\mathscr{H})$ with the poorer structure of symplectic manifold.

[12] R. Hermann, "Quantum mechanics and geometric analysis on manifolds," Int. J. Theor. Phys. **21**, 803 (1982).

[13] R. Cirelli, A. Manià, and L. Pizzocchero, "A smooth functional representation for C*-algebras," in preparation.

[14] This is the mathematically precise formulation of the variational principle appearing in any standard textbook on quantum mechanics: The eigenstates of a self-adjoint operator make stationary its mean value.

[15] B. Mielnik, "Generalized quantum mechanics," Commun. Math. Phys. **37**, 221 (1974).

[16] The fullness condition is linked with the homogeneity of the state manifold. At least in the finite-dimensional case, it is not difficult to show that, if $\mathscr{M}$ is connected, fullness implies local homogeneity: For every pair x, y of points of $\mathscr{M}$, there is a diffeomorphism between a neighborhood of x and a neighborhood of y that preserves ω, g and sends x into y. Homogeneity in the usual, global sense is granted under the additional assumption of completeness. As a partial converse, if $\mathscr{M}$ is connected, simply connected and (globally) homogeneous, then $\mathscr{K}(\mathscr{M}, \mathbf{R})$ is full.

[17] E. C. G. Stueckelberg, "Quantum theory in real Hilbert space," Helv. Phys. Acta **33**, 727 (1960); E. C. G. Stueckelberg and M. Guenin, "Quantum theory in real Hilbert space II (Addenda and Errata)," Helv, Phys. Acta **34**, 621 (1961).

[18] R. Cirelli, A. Manià, and L. Pizzocchero, J. Math. Phys. **31**, 2898 (1990).

Quantum mechanics as an infinite-dimensional Hamiltonian system with uncertainty structure: Part II

Renzo Cirelli, Alessandro Manià, and Livio Pizzocchero
Dipartimento di Fisica, Università di Milano, Milano, Italy and Istituto Nazionale di Fisica Nucleare, Sezione di Milano, Milano, Italy

(Received 22 August 1989; accepted for publication 27 June 1990)

Making reference to the formalism developed in Part I to formulate Schrödinger quantum mechanics, the properties of Kählerian functions in general, almost Kählerian manifolds, are studied.

I. INTRODUCTION

In a previous paper,[1] we have shown that a geometrical formulation is possible for Schrödinger quantum mechanics, in which the pure states of a quantum system are regarded as the points of an infinite-dimensional Kähler manifold and the observables are represented by a selected class of smooth functions on the state manifold (the *Kählerian functions*).

The central point in our work was the observation that a number of bilinear composition laws are defined naturally between smooth, real-, or complex-valued functions on any almost Kählerian manifold. Restricting these bilinear composition laws to Kählerian functions, we were able to reproduce in such a geometrical framework the algebraic structure of the observables of a quantum system, given by operator algebra in the conventional language of Schrödinger mechanics.

The Kähler manifold, on which we worked to obtain these results, is the projective space $\mathbf{P}(\mathscr{H})$ of the Hilbert space $\mathscr{H}$ of the system; we subsequently discussed the possibility of a generalization of Schrödinger mechanics in which the state manifold is not a projective space, and our conclusion was that no generalization of this kind is possible if one asks that observables are a sufficiently wide set and close an associative algebra.

Such a conclusion stems from a detailed analysis of the properties of Kählerian functions in general, almost Kählerian manifolds. This analysis, omitted in Ref. 1, is the object of the present paper. In Sec. II, we shall provide two characterizations of Kählerian functions; in Sec. III, we will show that the integrability of the almost complex structure is strictly connected with a condition of associativity on Kählerian functions; in Sec. IV, we will discuss the link between the holomorphic sectional curvature of the manifold and the closedness of the set of Kählerian functions with respect to $*_\nu$ product.

It is our opinion that these results have some interest even independently of the physical implications of Ref. 1. Indeed, from a purely mathematical viewpoint, they tell us how to associate to an almost Kählerian manifold a noncommutative algebra (the smooth functions with $*_\nu$ product) and how to recognize certain geometrical properties of the manifold analyzing a selected subspace of this algebra (the Kählerian functions).

II. CHARACTERIZATION OF KÄHLERIAN FUNCTIONS

Throughout this paper, in dealing with almost Kählerian manifolds, we shall maintain all the definitions and the notations introduced in Sec. 2 of Ref. 1. Furthermore, in agreement with Definition 4.3 of Ref. 1, we shall say that the set of the real Kählerian functions on an almost Kählerian manifold is *full* if the differentials of such functions at any point of the manifold span the whole cotangent space at that point. For the sake of brevity, we shall always omit the adjective "smooth" when speaking of tensor fields on the manifolds under examination; thus expressions such as "a vector field," "a one-form," etc. will stand for "a smooth vector field," "a smooth one-form," etc.

In this section, we will provide two characterizations (algebraic and geometrical) of the Kählerian functions on an arbitrary almost Kählerian manifold $(\mathscr{M}, J, g)$. The algebraic characterization is given by the following.

Proposition 2.1: For $f\in C^\infty(\mathscr{M},\mathbf{C})$, let $\mathscr{D}_f$ be the linear map

$$\mathscr{D}_f: C^\infty(\mathscr{M},\mathbf{C})\to C^\infty(\mathscr{M},\mathbf{C}), \quad l\mapsto \mathscr{D}_f l:=\{f,l\}.$$

The following statements are equivalent, for every $\nu\in\mathbf{R}\setminus\{0\}$: (a) f is Kählerian; (b) $\mathscr{D}_f$ is a derivation with respect to the Riemann bracket; (c) $\mathscr{D}_f$ is a derivation with respect to the Kähler bracket; (d) $\mathscr{D}_f$ is a derivation with respect to the $\circ_\nu$ product; and (e) $\mathscr{D}_f$ is a derivation with respect to the $*_\nu$ product.[2]

Proof: By Jacobi's identity, $\mathscr{D}_f$ is a derivation with respect to the Poisson bracket for every $f\in C^\infty(\mathscr{M},\mathbf{C})$. Using this fact, and keeping in mind that $<\,,\,>=((\,,\,))+i\{\,,\,\}$, we are easily led to the conclusion that (b) and (c) are equivalent. Furthermore, $\mathscr{D}_f$ is a derivation with respect to pointwise product, and thus (b) and (d), (c), and (e) are equivalent. It remains to show that (a)$\Leftrightarrow$(b); we can restrict our considerations to real functions.

Let $f, l, k\in C^\infty(\mathscr{M},\mathbf{R})$ and let X, Y, Z be the vector fields $I\,df$, $I\,dl$, $I\,dk$. Then,

$$\mathscr{D}_f((l,k)) = -L_X(g(Y,Z)) = -(L_X g)(Y,Z) - g(L_X Y,Z) - g(Y,L_X Z),$$

$$((\mathscr{D}_f l,k))=g(-L_X Y,Z), \quad ((l,\mathscr{D}_f k))=g(Y,-L_X Z),$$

and so $\mathscr{D}_f((l,k))$ equals $((\mathscr{D}_f k,l))+((k,\mathscr{D}_f l))$ if and only if

 0022-2488/90/122898-06$03.00

$$(L_X g)(Y,Z)=0.$$

This equation holds for arbitrary l, k if and only if $L_X g = 0$, i.e., if and only if f is Kählerian. ■

After the algebraic characterization of Kählerian functions provided by the previous proposition, we come to the geometric one. To obtain it, we must introduce the Riemannian connection on $\mathcal{M}$. This is the unique connection on $\mathcal{M}$ such that $\nabla g = 0$ and the torsion tensor is zero, i.e.,

$$\nabla_X Y - \nabla_Y X - [X,Y] = 0, \tag{2.1}$$

for every pair X, Y of vector fields. Of course, here we denote by ∇ the covariant derivative induced by the connection.[3] In particular, we shall work with the double covariant derivatives $\nabla\nabla f := \nabla(\nabla f)$ of smooth real-valued functions on $\mathcal{M}$, and with the covariant derivative ∇J of the almost complex structure. For $f \in C^\infty(\mathcal{M},\mathbf{R})$, $\nabla\nabla f$, as a bilinear form on $\mathcal{M}$, is symmetric:

$$(\nabla\nabla f)_x(X_x,Y_x) = (\nabla\nabla f)_x(Y_x,X_x), \tag{2.2}$$

for $x \in \mathcal{M}, X_x, Y_x \in T_x\mathcal{M}$. This is a well-known property of torsion-free connections. Regarding ∇J, it will be viewed as a section of $\mathcal{L}_2(T\mathcal{M},T\mathcal{M})$ (the vector bundle of base $\mathcal{M}$ whose fiber at any point x of $\mathcal{M}$ is the Banach space of bounded bilinear maps from $T_x\mathcal{M} \times T_x\mathcal{M}$ to $T_x\mathcal{M}$), in such a way that

$$(\nabla_Y J)X = \nabla J(X,Y), \tag{2.3}$$

for every pair X, Y of vector fields. It is known[4] that the almost complex structure is integrable if and only if $\nabla J = 0$.

We can now return to Kählerian functions. We start with the following lemma.

Lemma 2.2: Let $f \in C^\infty(\mathcal{M},\mathbf{R})$, and put $X := I\,df$. Then, for every pair Y, Z of vector fields,

$$\begin{aligned}(L_X g)(Y,Z) = {} & \nabla\nabla f(JY,Z) + \nabla\nabla f(JZ,Y) \\ & + \langle \nabla f, \nabla J(Y,Z) + \nabla J(Z,Y)\rangle.\end{aligned} \tag{2.4}$$

Proof: Let Y, Z be any pair of vector fields on $\mathcal{M}$. Then,

$$\begin{aligned}(L_X g)(Y,Z) &= L_X(g(Y,Z)) - g(L_X Y,Z) - g(Y,L_X Z) \\ &= \nabla_X(g(Y,Z)) - g(\nabla_X Y - \nabla_Y X, Z) \\ &\quad - g(Y,\nabla_X Z - \nabla_Z X),\end{aligned}$$

where we have used Eq. (2.1) to re-express $L_X Y = [X,Y]$ and $L_X Z = [X,Z]$. Since $\nabla_X(g(Y,Z)) = g(\nabla_X Y,Z) + g(Y,\nabla_X Z)$, from the previous equation it follows

$$(L_X g)(Y,Z) = g(\nabla_Y X,Z) + g(Y,\nabla_Z X). \tag{2.5}$$

We now remember that $X = I\,df = I\nabla f$. Then,

$$g(\nabla_Y X,Z) = g((\nabla_Y I)\nabla f, Z) + g(I(\nabla_Y \nabla f), Z).$$

We further make use of the identity

$$g(I\alpha,Z) = \langle \alpha, JZ\rangle, \tag{2.6}$$

holding for every one-form α and every vector field Z, from which follows taking the covariant derivative[5]

$$g((\nabla_Y I)\alpha,Z) = \langle \alpha, (\nabla_Y J)Z\rangle, \tag{2.7}$$

for every one-form α and every pair Z, Y of vector fields. Applying (2.6) and (2.7) with $\alpha = \nabla f$, we get

$$\begin{aligned}g(\nabla_Y X,Z) &= \langle \nabla f, (\nabla_Y J)Z\rangle + \langle \nabla_Y \nabla f, JZ\rangle \\ &= \langle \nabla f, \nabla J(Z,Y)\rangle + \nabla\nabla f(JZ,Y).\end{aligned}$$

In the same way, one treats the term $g(Y,\nabla_Z X)$ in Eq. (2.5). The final result is Eq. (2.4). ■

Proposition 2.3: Let $f \in C^\infty(\mathcal{M},\mathbf{R})$. Then, f is Kählerian if and only if for every pair Y, Z of vector fields

$$\begin{aligned}\nabla\nabla f(JY,JZ) = {} & \nabla\nabla f(Y,Z) - \langle \nabla f, \nabla J(Y,JZ) \\ & + \nabla J(JZ,Y)\rangle.\end{aligned} \tag{2.8}$$

Proof: Let $X := I\,df$; then f is Kählerian if and only if $L_X g = 0$. But $L_X g = 0$ if and only if $(L_X g)(Y,JZ) = 0$ for every pair Y, Z of vector fields. Writing down explicitly the equation $(L_X g)(Y,JZ) = 0$ with the help of Lemma 2.2, we obtain the wanted result. ■

The previous proposition provides the geometrical characterization of Kählerian functions; as we see, it puts a condition on the behavior of the "covariant Hessian" $\nabla\nabla f$ with respect to the almost complex structure J. Note that in the integrable case this condition simply means that $\nabla\nabla f$ is J invariant. On this we shall return later.

III. ASSOCIATIVITY AND INTEGRABILITY

In this section, we fix our attention on associators for triples of Kählerian functions with respect to the $*_\nu$ product. We show that the vanishing of such associators is strictly connected with the integrability of the almost complex structure J.

We start with a few identities about the covariant derivative of the almost complex structure in general, almost Kählerian manifolds.

Lemma 3.1: For every triple X, Y, Z of vector fields

(a) $\nabla J(JX,Y) + J\nabla J(X,Y) = 0$;

(b) $g(X,\nabla J(Y,Z)) + g(Y,\nabla J(X,Z)) = 0$;

(c) $g(X,\nabla J(JY,Z)) + g(Y,\nabla J(JX,Z)) = 0$.

Proof: (a) and (b) are obtained taking the covariant derivatives of the identities $J(JX) = -X$ and $g(X,JY) + g(Y,JX) = 0$. Furthermore, from (b) and the J invariance of g we get

$$g(X,\nabla J(JY,Z)) = -g(JY,\nabla J(X,Z)) = g(Y,J\nabla J(X,Z)).$$

By (a) this implies (c). ■

We now express $*_\nu$ associators for triples of Kählerian functions in geometrical terms.

Lemma 3.2: Let f, l, k be real Kählerian functions. Then, for every $\nu \in \mathbf{R}\setminus\{0\}$,

$$(4/\nu^2)((f *_\nu l) *_\nu k - f *_\nu (l *_\nu k))$$

$$= <<f,l>,k> - <f,<l,k>>$$

$$= -2\langle \nabla f, \nabla J(I\nabla k, G\nabla l)\rangle. \tag{3.1}$$

Proof: The first equality in (3.1) is obtained trivially for every triple of smooth functions and will hold in particular for real Kählerian functions. The rest of the proof will be devoted to the second equality.

From the very definition of Kähler bracket [Eq. (2.7) of Ref. 1], we compute

$$<<f,l>,k> - <f,<l,k>>$$

$$= [((\,(\,(f,l)),k)) - ((\,f,((l,k))\,)) - \{\{f,l\},k\} + \{f,\{l,k\}\}] + i[((\{f,l\},k)) + \{(\,(f,l)),k\} - ((\,f,\{l,k\})) - \{f,((l,k))\}].$$

Furthermore, using point (b) of Proposition 2.1, it is easily shown that the imaginary part of the associator vanishes. Hence,

$$<<f,l>,k> - <f,<l,k>>$$

$$= ((\,(\,(f,l)),k)) - ((\,f,((l,k))\,)) - \{\{f,l\},k\} + \{f,\{l,k\}\}. \tag{3.2}$$

This equation tells us that the associator of f, l and k with respect to $<\,,>$ is the difference of the associators with respect to $((\,,))$ and $\{\,,\}$. We now compute these two associators.

By definition

$$((\,(\,f,l)),k)) = \langle \nabla\langle \nabla f, G\nabla l\rangle, G\nabla k\rangle.$$

The covariant derivative $\nabla\langle \nabla f, G\nabla l\rangle$ can be computed using the properties of ∇ and the identity $\langle \alpha, G\beta\rangle = \langle \beta, G\alpha\rangle$, holding for every pair α, β of one forms on $\mathcal{M}$; the result is

$$\langle \nabla\langle \nabla f, G\nabla l\rangle, X\rangle = \nabla\nabla f(G\nabla l, X) + \nabla\nabla l(G\nabla f, X),$$

for every vector field X. In particular, for $X = G\nabla k$, we obtain

$$((\,(\,(f,l)),k)) = \nabla\nabla f(G\nabla l, G\nabla k) + \nabla\nabla l(G\nabla f, G\nabla k).$$

From this, we also obtain, interchanging the roles of f and k, a similar expression for $((\,f,((l,k))\,))$. Subtracting, we get

$$((\,(\,(f,l)),k)) - ((\,f,((l,k))\,))$$

$$= \nabla\nabla f(G\nabla l, G\nabla k) - \nabla\nabla k(G\nabla l, G\nabla f). \tag{3.3}$$

Now let us pass to the Poisson bracket. From the definition of $\{\,,\}$ and the identity $I = -J\circ G$,

$$\{\{f,l\},k\} = \langle \nabla\langle \nabla f, JG\nabla l\rangle, JG\nabla k\rangle.$$

By a slight modification of the technique employed to compute $((\,(\,(f,l)),k))$, we get

$$\{\{f,l\},k\} = \nabla\nabla f(JG\nabla l, JG\nabla k) - \nabla\nabla l(JG\nabla f, JG\nabla k) + \langle \nabla f, \nabla J(G\nabla l, JG\nabla k)\rangle.$$

This also allows us to express $\{f,\{l,k\}\}$ and thus, subtracting

$$\{\{f,l\},k\} - \{f,\{l,k\}\}$$

$$= \nabla\nabla f(JG\nabla l, JG\nabla k) - \nabla\nabla k(JG\nabla l, JG\nabla f\,) + \langle \nabla f, \nabla J(G\nabla l, JG\nabla k)\rangle - \langle \nabla k, \nabla J(G\nabla l, JG\nabla f\,)\rangle.$$

From Eqs. (3.2)–(3.4), it follows

$$<<f,l>,k> - <f,<l,k>>$$

$$= -(\nabla\nabla f(JG\nabla l, JG\nabla k) - \nabla\nabla f(G\nabla l, G\nabla k)) - \langle \nabla f, \nabla J(G\nabla l, JG\nabla k)\rangle + (\nabla\nabla k(JG\nabla l, JG\nabla f\,) - \nabla\nabla k(G\nabla l, G\nabla f\,)) + \langle \nabla k, \nabla J(G\nabla l, JG\nabla f\,)\rangle.$$

Now, let us use for f and k the characterization of Kählerian functions provided by Proposition 2.3; in this way we obtain

$$<<f,l>,k> - <f,<l,k>>$$

$$= \langle \nabla f, \nabla J(JG\nabla k, G\nabla l)\rangle - \langle \nabla k, \nabla J(JG\nabla f, G\nabla l)\rangle.$$

We finally apply point (c) of Lemma 3.1 with the substitutions $X = G\nabla k$, $Y = G\nabla f$, $Z = G\nabla l$. Inserting the result in the previous equation and reexpressing JG as $-I$, we obtain the second equality in (3.1). ∎

Proposition 3.3: Let $\nu \in \mathbf{R}\setminus\{0\}$ be arbitrarily chosen. If the almost complex structure J is integrable, for every triple f, l, k of complex valued Kählerian functions the associator $(f *_\nu l) *_\nu k - f *_\nu (l *_\nu k)$ vanishes.

As a partial converse, if $\mathcal{K}(\mathcal{M},\mathbf{R})$ is full and the associators of all triples of Kählerian functions vanish, then J is integrable.

Proof: By linearity arguments, if suffices to consider the associators for real Kählerian functions. If $\nabla J = 0$, by Lemma 3.2 the $*_\nu$ associators of all triples of real Kählerian functions vanish. Conversely, the vanishing of such associators means

$$\langle \nabla f, \nabla J(I\nabla k, G\nabla l)\rangle = 0, \quad \forall f,k,l \in \mathcal{K}(\mathcal{M},\mathbf{R}).$$

In the case $\mathcal{K}(\mathcal{M},\mathbf{R})$ is full, this implies $\nabla J = 0$. ∎

We point out that Proposition 3.3 does not mean that $\mathcal{K}(\mathcal{M},\mathbf{C})$ with $*_\nu$ product is an associative algebra when J is integrable. Indeed, in general $*_\nu$ product does not close in $\mathcal{K}(\mathcal{M},\mathbf{C})$. Necessary and sufficient conditions for closedness will be provided in the next section, under the assumption of integrability.

IV. ALGEBRAIC CLOSEDNESS AND CURVATURE

Throughout this section, the almost complex structure J will be assumed integrable; thus $(\mathcal{M},J,g)$ is a Kähler manifold. Our purpose is to discuss the connection between the closedness of $\mathcal{K}(\mathcal{M},\mathbf{C})$ with respect to $*_\nu$ product and

the constancy of the holomorphic sectional curvature. It is convenient to recall briefly all the machinery of curvature in Kähler manifolds.

We shall denote by R the curvature tensor of the metric connection; R can be viewed as a section of the tensor bundle $\mathscr{L}_3(T\mathscr{M},T\mathscr{M})$ (see Ref. 6). Rather than with R, we shall work with its completely covariant version K. This is the quadrilinear form on $\mathscr{M}$ defined by

$$K_x(X_x,Y_x,Z_x,W_x)=g_x(X_x,R_x(Y_x,Z_x,W_x)), \tag{4.1}$$

for $x\in\mathscr{M}, X_x,Y_x,Z_x,W_x\in T_x\mathscr{M}$. The defect in commutativity in the second covariant derivative of any one-form α is expressed through K as follows:

$$\nabla\nabla\alpha(X,Y,Z)=\nabla\nabla\alpha(X,Z,Y)+K(X,G\alpha,Z,Y) \tag{4.2}$$

for every triple X, Y, Z of vector fields.

The tensor K has some remarkable properties of symmetry and invariance that we quote here: For $x\in\mathscr{M}, X_x,Y_x,Z_x,W_x\in T_x\mathscr{M}$.

$$\begin{aligned}K_x(X_x,Y_x,Z_x,W_x)&=-K_x(Y_x,X_x,Z_x,W_x)\\&=-K_x(X_x,Y_x,W_x,Z_x)=K_x(Z_x,W_x,X_x,Y_x);\end{aligned} \tag{4.3}$$

$$K_x(X_x,Y_x,Z_x,W_x)+K_x(X_x,Z_x,W_x,Y_x)+K_x(X_x,W_x,Y_x,Z_x)=0; \tag{4.4}$$

$$\begin{aligned}&K_x(J_xX_x,J_xY_x,Z_x,W_x)\\&=K_x(X_x,Y_x,J_xZ_x,J_xW_x)\\&=K_x(X_x,Y_x,Z_x,W_x).\end{aligned} \tag{4.5}$$

Finally, if $x\in\mathscr{M}$ and π_x is a J_x-invariant two-plane of $T_x\mathscr{M}$, the holomorphic sectional curvature $c(\pi_x)$ is defined by

$$c(\pi_x):=K_x(U_x,V_x,U_x,V_x), \tag{4.6}$$

where (U_x,V_x) is any orthonormal basis of π_x. The Kähler manifold $(\mathscr{M},J,g)$ is said to be of constant holomorphic sectional curvature $c(c\in\mathbf{R})$ if

$$c(\pi_x)=c,$$

for every $x\in\mathscr{M}$ and every J_x-invariant two-plane π_x of $T_x\mathscr{M}$. This property is satisfied if and only if $K=cH$, where H is the quadrilinear form on $\mathscr{M}$ defined by

$$\begin{aligned}&H_x(X_x,Y_x,Z_x,W_x)\\&:=\tfrac{1}{4}\{g_x(X_x,Z_x)g_x(Y_x,W_x)-g_x(X_x,W_x)g_x(Y_x,Z_x)\\&\quad+g_x(J_xX_x,Z_x)g_x(J_xY_x,W_x)-g_x(J_xX_x,W_x)\\&\quad\times g_x(J_xY_x,Z_x)\\&\quad+2g_x(J_xX_x,Y_x)g(J_xZ_x,W_x)\}.\end{aligned} \tag{4.7}$$

We now come to closedness of $\mathscr{K}(\mathscr{M},\mathbf{C})$ with respect to the $*_\nu$ product for a fixed $\nu\in\mathbf{R}\setminus\{0\}$. Since J is integrable, the characterization of real Kählerian functions provided by Proposition 2.3 becomes rather simple and effective. Equivalent formulations are

$$\nabla\nabla f(JX,JY)=\nabla\nabla f(X,Y) \tag{4.8}$$

for every pair X, Y of vector fields, or

$$\mathscr{K}f=0, \tag{4.9}$$

where $\mathscr{K}:C^\infty(\mathscr{M},\mathbf{R})\to\mathrm{Sec}(\mathscr{L}_2(T\mathscr{M},\mathbf{R}))$ is the second-order differential operator defined by

$$\begin{aligned}(\mathscr{K}f)_x(X_x,Y_x)&=(\nabla\nabla f)_x(J_xX_x,J_xY_x)\\&\quad-(\nabla\nabla f)_x(X_x,Y_x)\end{aligned} \tag{4.10}$$

for $x\in\mathscr{M}, X_x,Y_x\in T_x\mathscr{M}$.

We now remember[1] that the closedness of $\mathscr{K}(\mathscr{M},\mathbf{C})$ with respect to the $*_\nu$ product is equivalent to the closedness of $\mathscr{K}(\mathscr{M},\mathbf{R})$ with respect to the $\circ_\nu$ product. Let f, l be real Kählerian functions; then $f\circ_\nu l$ is Kählerian if and only if $\mathscr{K}(f\circ_\nu l)=0$. In order to evaluate $\mathscr{K}(f\circ_\nu l)$, we start from the computation of $\mathscr{K}((f,l))$; subsequently, we shall compute $\mathscr{K}(fl)$.

Lemma 4.1: For $f,l\in\mathscr{K}(\mathscr{M},\mathbf{R})$,

$$\begin{aligned}&\mathscr{K}((f,l))(X,Y)\\&=-K(G\nabla f,X,G\nabla l,Y)+K(G\nabla f,JX,G\nabla l,JY)\\&\quad-K(G\nabla l,X,G\nabla f,Y)+K(G\nabla l,JX,G\nabla f,JY),\end{aligned} \tag{4.11}$$

for every pair X, Y of vector fields.

Proof: Let us compute $\nabla\nabla((f,l))=\nabla\nabla\langle\nabla f,G\nabla l\rangle$. By standard manipulations we get

$$\begin{aligned}\nabla\nabla((f,l))(X,Y)&=\nabla\nabla\nabla f(G\nabla l,X,Y)+\nabla\nabla f(G(\nabla_Y\nabla l),X)\\&\quad+\nabla\nabla\nabla l(G\nabla f,X,Y)\\&\quad+\nabla\nabla l(G(\nabla_Y\nabla f),X).\end{aligned} \tag{4.12}$$

Now, from the J-invariance of $\nabla\nabla l$ and g

$$G(\nabla_{JY}\nabla l)=-G((\nabla_Y\nabla l)\circ J)=JG(\nabla_Y\nabla l).$$

From here and from the J-invariance of $\nabla\nabla f$,

$$\begin{aligned}\nabla\nabla f(G(\nabla_{JY}\nabla l),JX)&=\nabla\nabla f(JG(\nabla_Y\nabla l),JX)\\&=\nabla\nabla f(G(\nabla_Y\nabla l),X).\end{aligned}$$

In the same way one shows that the term $\nabla\nabla l(G(\nabla_Y\nabla f),X)$ in Eq. (4.12) is invariant under the substitution $X,Y\mapsto JX,JY$. Therefore,

$$\begin{aligned}&\nabla\nabla((f,l))(JX,JY)-\nabla\nabla((f,l))(X,Y)\\&=\nabla\nabla\nabla f(G\nabla l,JX,JY)-\nabla\nabla\nabla f(G\nabla l,X,Y)\\&\quad+\nabla\nabla\nabla l(G\nabla f,JX,JY)-\nabla\nabla\nabla l(G\nabla f,X,Y).\end{aligned} \tag{4.13}$$

We now observe that from

$$\nabla\nabla f(JX,JY)=\nabla\nabla f(X,Y),$$

it follows, taking the covariant derivative,

$$\nabla\nabla\nabla f(JX,JY,Z)=\nabla\nabla\nabla f(X,Y,Z), \tag{4.14}$$

for any triple X, Y, Z of vector fields. Furthermore, applying Eq. (4.2) with $\alpha=\nabla f$, we obtain

$$\nabla\nabla\nabla f(X,Y,Z)=\nabla\nabla\nabla f(X,Z,Y)+K(X,G\nabla f,Z,Y),$$

or, using the symmetry of $\nabla\nabla\nabla f$ in the first two arguments (that follows from the symmetry of $\nabla\nabla f$)

$$\nabla\nabla\nabla f(X,Y,Z)=\nabla\nabla\nabla f(Z,X,Y)+K(X,G\nabla f,Z,Y). \tag{4.15}$$

Equations (4.14) and (4.15) together give

$$\nabla\nabla\nabla f(Z,JX,JY)=\nabla\nabla\nabla f(Z,X,Y)+K(X,G\nabla f,Z,Y)-K(JX,G\nabla f,Z,JY),$$

that is, using Eq. (4.3),

$$\nabla\nabla\nabla f(Z,JX,JY)=\nabla\nabla\nabla f(Z,X,Y)-K(G\nabla f,X,Z,Y)+K(G\nabla f,JX,Z,JY). \tag{4.16}$$

From Eqs. (4.13) and (4.16) the thesis is easily inferred. To get it, it suffices to insert in (4.13) the equation obtained specializing (4.16) to the case $Z=G\nabla l$, and its analogous with f and l interchanged. ■

Lemma 4.2: For f, l in $\mathscr{K}(\mathscr{M},\mathbf{R})$,

$$\begin{aligned}\mathscr{K}(fl)(X,Y)&=H(G\nabla f,X,G\nabla l,Y)-H(G\nabla f,JX,G\nabla l,JY)\\&\quad+H(G\nabla l,X,G\nabla f,Y)-H(G\nabla l,JX,G\nabla f,JY),\end{aligned} \tag{4.17}$$

for every pair X, Y of vector fields, where H is the quadrilinear form on $\mathscr{M}$ defined in Eq. (4.7).

Proof: For every pair X, Y of vector fields,

$$\nabla\nabla(fl)(X,Y)=f\nabla\nabla l(X,Y)+l\nabla\nabla f(X,Y)+\langle\nabla f,X\rangle\langle\nabla l,Y\rangle+\langle\nabla l,X\rangle\langle\nabla f,Y\rangle.$$

From here and from the J-invariance of $\nabla\nabla f$, $\nabla\nabla l$ it follows

$$\mathscr{K}(fl)(X,Y)=\langle\nabla f,JX\rangle\langle\nabla l,JY\rangle+\langle\nabla l,JX\rangle\langle\nabla f,JY\rangle-\langle\nabla f,X\rangle\langle\nabla l,Y\rangle-\langle\nabla l,X\rangle\langle\nabla f,Y\rangle. \tag{4.18}$$

But, from the explicit expression of H, it follows, after some manipulations,

$$\begin{aligned}&H(G\nabla f,X,G\nabla l,Y)-H(G\nabla f,JX,G\nabla l,JY)\\&\quad=-\tfrac{1}{2}\langle\nabla f,Y\rangle\langle\nabla l,X\rangle+\tfrac{1}{2}\langle\nabla f,JY\rangle\langle\nabla l,JX\rangle\\&\qquad+\tfrac{1}{2}\langle\nabla f,JX\rangle\langle\nabla l,JY\rangle-\tfrac{1}{2}\langle\nabla f,X\rangle\langle\nabla l,Y\rangle.\end{aligned} \tag{4.19}$$

Summing to Eq. (4.19) the analogous one obtained interchanging the roles of f and l, and comparing with (4.18), we obtain the wanted result. ■

Proposition 4.3: If the Kähler manifold $(\mathscr{M},J,g)$ has constant holomorphic section curvature $2/\nu$, ($\nu\in\mathbf{R}\setminus\{0\}$), then $\mathscr{K}(\mathscr{M},\mathbf{C})$ is closed with respect to $*_\nu$ product.

As a partial converse, if $\mathscr{K}(\mathscr{M},\mathbf{R})$ is full and $\mathscr{K}(\mathscr{M},\mathbf{C})$ is closed with respect to $*_\nu$ product, then $(\mathscr{M},J,g)$ has constant holomorphic sectional curvature $2/\nu$.

Proof: We shall use the fact, already outlined, that the closedness of $\mathscr{K}(\mathscr{M},\mathbf{C})$ with respect to $*_\nu$ product is equivalent to the closedness of $\mathscr{K}(\mathscr{M},\mathbf{R})$ with respect to $\circ_\nu$ product.

Assume $(\mathscr{M},J,g)$ has constant curvature $2/\nu$. Then, $K=(1/\nu)H$ and thus, comparing Eqs. (4.11) and (4.17), we obtain that for every pair f, l of real Kählerian functions one has

$$\mathscr{K}((f,l))=-(2/\nu)\,\mathscr{K}(fl) \tag{4.20}$$

that is $\mathscr{K}(\frac{1}{2}\nu((f,l))+fl)=0$. This tells us that the product $f\circ_\nu l$ is again a Kählerian function.

Conversely, if $\mathscr{K}(\mathscr{M},\mathbf{R})$ is closed with respect to $\circ_\nu$ product, for every pair f,l of real Kählerian functions Eq. (4.20) holds. Reexpressing $\mathscr{K}((f,l))$ and $\mathscr{K}(fl)$ via Eqs. (4.11) and (4.17) and using the fullness assumption we are easily led to the conclusion that for $x\in\mathscr{M}$ and $Z_x,X_x,W_x,Y_x\in T_x\mathscr{M}$,

$$\begin{aligned}&-K_x(Z_x,X_x,W_x,Y_x)+K_x(Z_x,J_xX_x,W_x,J_xY_x)-K_x(W_x,X_x,Z_x,Y_x)+K_x(W_x,J_xX_x,Z_x,J_xY_x)\\&\quad=-(2/\nu)H_x(Z_x,X_x,W_x,Y_x)+(2/\nu)H_x(Z_x,J_xX_x,W_x,J_xY_x)\\&\qquad-(2/\nu)H_x(W_x,X_x,Z_x,Y_x)+(2/\nu)H_x(W_x,J_xX_x,Z_x,J_xY_x).\end{aligned} \tag{4.21}$$

We want to show that this equation implies $K_x=(2/\nu)H_x$. Indeed, let us apply (4.21) with the substitutions $X_x=J_xZ_x$, $W_x=Z_x$, $Y_x=J_xZ_x$. We obtain

$$-2K_x(Z_x,J_xZ_x,Z_x,J_xZ_x)\doteq-(4/\nu)H_x(Z_x,J_xZ_x,Z_x,J_xZ_x). \tag{4.22}$$

But K_x and $(2/\nu)H_x$ both satisfy Eqs. (4.3)–(4.5); therefore, as shown in Ref. 4, Chap. IX, Proposition 7.1, Eq. (4.22) implies $K_x=(2/\nu)H_x$. ■

We conclude with a few remarks. Proposition 4.3 admits a limiting version with $\nu=\infty$, if one replaces $*_\nu$ and $\circ_\nu$ product with Kähler and Riemann bracket, respectively [this is natural, since in the definitions of $*_\nu$ and $\circ_\nu$ the terms containing $<,>$ and $((,))$ are dominating as ν grows to infinity]. The $\nu=\infty$ statement is easily established by a trivial modification of the proof of 4.3, and sounds as follows: If $(\mathscr{M},J,g)$ has constant sectional holomorphic sectional curvature 0, then $\mathscr{K}(\mathscr{M},\mathbf{C})$ is closed

with respect to $<\ ,\ >$; if $\mathscr{K}(\mathscr{M},\mathbf{R})$ is full and $\mathscr{K}(\mathscr{M},\mathbf{C})$ is closed with respect to $<\ ,\ >$, then $(\mathscr{M},J,g)$ has constant holomorphic sectional curvature 0.

Obviously enough, the standard model for a Kähler manifold of zero curvature is a complex Hilbert space with (the real part of) the scalar product regarded as a metric; the standard model for the negative curvature case is also known.[4]

[1] R. Cirelli, A. Manià, and L. Pizzocchero, J. Math. Phys. **31**, 2891 (1990).

[2] In Ref. 1, we were prevalently interested in the case $\nu > 0$ because of applications to quantum mechanics where $\nu = \hbar$. In the present framework, which is a bit more abstract, we shall also admit the possibility that ν is negative. The case $\nu = 0$, in which both $\circ_\nu$ and $*_\nu$ product become pointwise product, is of a particular nature; for example, for $\nu = 0$ points (d) and (e) in Proposition 2.1. are not equivalent to (a), (b), and (c).

[3] The theory of connections in vector bundles is well known in the finite-dimensional case. A suitable formulation is possible also in the Banachic case (or in the Hilbertian case, for Riemannian connections), in which all the results needed here can be generalized from finite dimensions.

[4] S. Kobayashi and K. Nomizu, *Foundations of Differential Geometry* (Wiley, New York, 1969), Vols. I and II.

[5] In order to avoid misunderstandings, we explain in which sense the identity (2.7) is the covariant derivative of the identity (2.6). Equation (2.6) tells us that $A = B$, where A and B are the sections of $\mathscr{L}_2(T^*\mathscr{M} \times T\mathscr{M},\mathbf{R})$ such that $A(\alpha,Z) = g(I\alpha,Z)$ and $B(\alpha,Z) = \langle \alpha, JZ \rangle$ for every one form α and every vector field Z. From $A = B$ it follows $\nabla_Y A = \nabla_Y B$ for every vector field Y; but, for every pair Z, Y of vector fields and every one-form α, $(\nabla_Y A)(\alpha,Z)$ and $(\nabla_Y B) \times (\alpha,Z)$ are just the left and the right side of (2.7). Similar manipulations will be frequently used in this paper, without any further comments.

[6] R. S. Hamilton, "The inverse function theorem of Nash and Moser," Bull. Am. Math. Soc. 7, 65 (1982). The theory of infinite-dimensional connections proposed in this reference is suitable for a more general framework than the one in which we work here (indeed, it is formulated in the context of Fréchet manifolds and Fréchet vector bundles). We quote this paper only to give an indication on how to define the curvature tensor.

Bäcklund transformation, conservation laws, and inverse scattering transform of a model integrodifferential equation for water waves

Y. Matsuno
Department of Physics, Faculty of Liberal Arts, Yamaguchi University, Yamaguchi 753, Japan

(Received 20 February 1990; accepted for publication 18 July 1990)

The Bäcklund transformation (BT), an infinite number of conservation laws, and the inverse scattering transform (IST) of a model integrodifferential equation for water waves in fluids of finite depth [Y. Matsuno, J. Math. Phys. **29**, 49(1989)] are constructed by employing the bilinear transformation method. The model equation is also shown to pass the Painlevé test. These facts prove the complete integrability of the equation. Both the deep- and shallow-water limits of various results thus obtained are then investigated in detail. In addition, a new method to evaluate conserved quantities for pure N-soliton is developed by utilizing actively the time part of the BT. It is found that the structure of conservation laws exhibits peculiar characteristics in comparison with those of usual water wave equations such as the Benjamin–Ono and the Korteweg–de Vries equations. The most important problem left open in this paper is to solve various IST equations.

I. INTRODUCTION

In this paper, we consider the following integrodifferential evolution equation:

$$u_t + u_x - 2uu_t + 2u_x \int_x^\infty u_t\,dx - Tu_{tx} = 0, \quad u = u(x,t), \tag{1.1a}$$

where the operator T is defined by

$$Tu(x,t) = \frac{1}{2\delta} P \int_{-\infty}^{\infty} \left\{ \coth\left[\frac{\pi(y-x)}{2\delta}\right] - \operatorname{sgn}(y-x) \right\} u(y,t)\,dy, \tag{1.1b}$$

and the subscripts t and x appended to u denote partial differentiation.

Equation (1.1) has already been proposed as a model equation describing wave phenomena in fluids of finite depth.[1] The parameter δ in (1.1b) represents the depth of fluids. In the deep-water limit $\delta \to \infty$, it reduces to the equation

$$u_t + u_x - 2uu_t + 2u_x \int_x^\infty u_t\,dx - Hu_{tx} = 0, \tag{1.2a}$$

where H is the Hilbert transform given by

$$Hu(x,t) = \frac{1}{\pi} P \int_{-\infty}^{\infty} \frac{u(y,t)}{y-x}\,dy. \tag{1.2b}$$

On the other hand, in the shallow-water limit $\delta \to 0$, if we introduce the new variables τ and ξ by the relations

$$\tau = \sqrt{3/\delta}\,t, \tag{1.3a}$$

$$\xi = \sqrt{3/\delta}\,x, \tag{1.3b}$$

and use the expansion formula (B8) in Appendix B, Eq. (1.1) reduces to the shallow-water wave equation that has already been proposed by Hirota and Satsuma[2]

$$u_\tau + u_\xi - 2uu_\tau + 2u_\xi \int_\xi^\infty u_\tau\,d\xi - u_{\tau\xi\xi} = 0. \tag{1.4}$$

Therefore, Eq. (1.1) is an intermediate version between Eqs. (1.2) and (1.4). In this respect, it should be remarked that an evolution equation of the form[3,4]

$$u_t + u_x + 2uu_x + Tu_{xx} = 0 \tag{1.5}$$

is known as an intermediate equation between the Benjamin–Ono (BO)[5,6] and the Korteweg–de Vries (KdV) equations. The mathematical structure of Eq. (1.5) is now completely understood.[7–9] However, the characteristics of Eq. (1.1) are quite different from those of Eq. (1.5) as shown in this paper.

The multisoliton solutions[1,2] of Eqs. (1.1), (1,2), and (1.4) have already been obtained by employing the bilinear transformation method.[10–12] In particular, the properties of solutions of Eq. (1.2) have been studied in great detail[13] and it was found that it exhibits an algebraic N-soliton solution expressed in terms of Pfaffians.[14]

The purpose of the present paper is to construct the Bäcklund transformation (BT), an infinite number of conservation laws and the inverse scattering transform (IST) of Eq. (1.1) and to prove its complete integrability. Both the deep- and shallow-water limits are then taken for various results thus obtained and which give rise to the corresponding ones for Eqs. (1.2) and (1.4), respectively. Throughout this paper, we mainly use the bilinear transformation method[10–12] as a mathematical tool.

In Sec. II, the BT, an infinite number of conservation laws, and the IST are constructed for Eq. (1.1). It is also demonstrated that Eq. (1.1) passes the so-called Painlevé test.[15–18] In Sec. III, the deep-water limit of the results obtained in Sec. II is investigated in detail. In addition to this, the conserved quantities are evaluated explicitly for pure N-soliton solution and an initial condition evolving into pure N solitons is briefly discussed. In Sec. IV, the shallow-water limit is considered in the same way. Section V is devoted to concluding remarks where a few comments are made concerning problems left open in this paper. In Appendix A, the formulas of bilinear operators are noted for the convenience of the reader unfamiliar with the bilinear formalism. Finally,

 0022-2488/90/122904-13$03.00

in Appendix B, various properties of the singular integral operators T and H are described.

II. STUDY OF EQ. (1.1)

First of all, let us bilinearize Eq. (1.1). For the purpose, we introduce the following dependent variable transformation:

$$u=\frac{i}{2}\frac{\partial}{\partial x}\ln\frac{f_+}{f_-}, \tag{2.1a}$$

with

$$f_+(x,t)=f(x-i\delta,t), \tag{2.1b}$$

$$f_-(x,t)=f(x+i\delta,t), \tag{2.1c}$$

where $f(z,t)$ is an analytic function of z and it is assumed that $f(z-i\delta,t)$ has no zero in the complex region $0\leqslant \mathrm{Im}\, z\leqslant 2\delta$. It then follows by using the Cauchy residue theorem that

$$Tu_x=-\frac{1}{2}\frac{\partial^2}{\partial x^2}(f_+f_-)+\delta^{-1}u. \tag{2.2}$$

Substituting (2.1a) and (2.2) into Eq. (1.1) and integrating once with respect to x, we obtain the following bilinear equation for f_+ and f_-:[1]

$$[i(1-\delta^{-1})D_t+iD_x+D_tD_x]f_+\cdot f_-=0. \tag{2.3}$$

Here, the bilinear operators D_t and D_x are defined by the relation

$$D_t^mD_x^nf_+\cdot f_-=\left(\frac{\partial}{\partial t}-\frac{\partial}{\partial t'}\right)^m\left(\frac{\partial}{\partial x}-\frac{\partial}{\partial x'}\right)^nf_+(x,t)\times f_-(x',t')\Big|_{\substack{t'=t\\ x'=x}},\quad (m,n=0,1,\ldots). \tag{2.4}$$

Equation (2.3) is the basic equation and plays a fundamental role in this section.

A. Bäcklund transformation

In this subsection, starting from the bilinear equation (2.3) we construct the BT in bilinear form and then transform it into ordinary form.

1. BT in bilinear form

We now construct the BT of Eq. (1.1) in bilinear form. Let another solution of Eq. (2.3) be $g_\pm$, namely,

$$[i(1-\delta^{-1})D_t+iD_x+D_tD_x]g_+\cdot g_-=0, \tag{2.5}$$

and consider the following equation:

$$P\equiv g_+g_-[i(1-\delta^{-1})D_t+iD_x+D_tD_x]f_+\cdot f_- -\{[i(1-\delta^{-1})D_t+iD_x+D_tD_x]\times g_+\cdot g_-\}f_+f_-=0. \tag{2.6}$$

Obviously, if $f_\pm$ is a solution of Eq. (2.3), then $g_\pm$ gives an another solution and vice versa. We prove that Eq. (2.6) is satisfied identically by the following system of bilinear equations that relate $f_\pm$ and $g_\pm$ and hence these equations constitute a BT of Eq. (2.3):

$$D_tg_+\cdot f_- -D_tg_-\cdot f_+=-i(g_+f_- +g_-f_+)+\nu(g_+f_- -g_-f_+), \tag{2.7a}$$

$$D_xg_+\cdot f_- +D_xg_-\cdot f_+=-i(1-\delta^{-1})\times(g_+f_- -g_-f_+)+\mu(g_+f_- +g_-f_+), \tag{2.7b}$$

$$D_tD_xg_+\cdot f_+=\mu D_tg_+\cdot f_+ +\nu D_xg_+\cdot f_+ +\kappa f_+g_+, \tag{2.7c}$$

$$D_tD_xg_-\cdot f_-=\mu D_tg_-\cdot f_- +\nu D_xg_-\cdot f_- +\kappa g_-f_-. \tag{2.7d}$$

Here, μ, ν, and κ are arbitrary parameters that may be related to each other by specifying boundary conditions. To show the above BT, we first use (A3) and (A4) to modify (2.6) into the form

$$P=D_t\{g_-f_+\cdot[D_xg_+\cdot f_- +i(1-\delta^{-1})g_+f_-]\} +D_x[g_-f_+\cdot(D_tg_+\cdot f_- +ig_+f_-)] +(D_tD_xg_+\cdot f_+)g_-f_- -g_+f_+D_tD_xg_-\cdot f_-. \tag{2.8a}$$

Substituting (2.7c) and (2.7d) into (2.8a) and using (A3), P is transformed into

$$P=D_t\{g_-f_+\cdot[D_xg_+\cdot f_- +(i(1-\delta^{-1})-\mu)g_+f_-]\} +D_x\{g_-f_+\cdot[D_tg_+\cdot f_- +(i-\nu)g_+f_-]\}. \tag{2.8b}$$

Finally, introducing (2.7a) and (2.7b) into (2.8b) and using (A1) and (A2), P becomes

$$P=-D_t(g_-f_+\cdot D_xg_-\cdot f_+)+D_x(g_-f_+\cdot D_tg_-\cdot f_+)=0. \tag{2.8c}$$

This completes the proof of Eq. (2.6).

2. BT in ordinary form

In order to rewrite the BT in ordinary form, it is convenient to introduce the potentials $\bar{u}$ and $\bar{v}$ by

$$\bar{u}=(i/2)\ln(f_+/f_-), \tag{2.9a}$$

$$\bar{v}=(i/2)\ln(g_+/g_-). \tag{2.9b}$$

It then follows from (2.2) and (2.9) that

$$f_\pm=f_{\pm,0}\left[\mp 2i\hat{P}_\mp(\bar{u}-\bar{u}_0)+\delta^{-1}\int_{-\infty}^{x}(\bar{u}-\bar{u}_0)dx\right], \tag{2.10a}$$

$$g_\pm=g_{\pm,0}\left[\mp 2i\hat{P}_\mp(\bar{v}-\bar{v}_0)+\delta^{-1}\int_{-\infty}^{x}(\bar{v}-\bar{v}_0)dx\right]. \tag{2.10b}$$

Here, $P_\pm$ are integral operators defined by

$$\hat{P}_\pm=\tfrac{1}{2}(1\pm iT), \tag{2.11}$$

and $\bar{u}_0=\bar{u}(-\infty,t)$, $\bar{v}_0=\bar{v}(-\infty,t)$, $f_{\pm,0}=f_\pm(-\infty,t)$, $g_{\pm,0}=g_\pm(-\infty,t)$. If we subtract (2.7d) from (2.7c) after substituting (2.10) into these equations and define the function w by

$$w=\bar{v}-\bar{u}, \tag{2.12}$$

with the boundary conditions

$$w(+\infty,t) = w(-\infty,t) = w_0, \tag{2.13}$$

then, we obtain the time part of the BT as follows:

$$w_{tx} - w_t T w_x - w_x T w_t + \delta^{-1} w_x \int_{-\infty}^{x} w_t\, dx + \delta^{-1}(w - w_0) w_t = \mu w_t + \nu w_x - 2\bar{u}_{tx}. \tag{2.14}$$

Integrating (2.14) and using (B5), we obtain an important relation

$$\int_{-\infty}^{\infty} w_t\, dx = 0. \tag{2.15}$$

This fact will be used in Sec. II B to derive conservation laws.

The space part of the BT stems by substituting (2.10) into (2.7b) and the result is expressed in the form

$$e^{2iw} = \frac{2i\hat{P}_- w_x - \delta^{-1}(w - w_0) + 2iu - i(1-\delta^{-1}) + \mu}{2i\hat{P}_+ w_x + \delta^{-1}(w - w_0) + 2iu - i(1-\delta^{-1}) - \mu}. \tag{2.16}$$

Similarly, Eq. (2.7a) can be rewritten as

$$e^{2iw} = -\frac{2i\hat{P}_- w_t - \delta^{-1}\int_{-\infty}^{x} w_t\, dx + 2i\bar{u}_t - i + \nu}{2i\hat{P}_+ w_t + \delta^{-1}\int_{-\infty}^{x} w_t\, dx + 2i\bar{u}_t - i - \nu}. \tag{2.17}$$

Equation (2.17) may be regarded as a time part of the BT in place of Eq. (2.14). Now, by taking $|x| \to \infty$ in (2.17) and using (2.13) with the boundary conditions $u(\pm\infty,t) = w_x(\pm\infty,t) = 0$, one sees that w_0 is expressed in terms of ν as

$$w_0 = (1/2i)\ln[(\nu - i)/(\nu - i)]. \tag{2.18}$$

Furthermore, if we equate (2.16) and (2.17) and then take $|x| \to \infty$, we find that μ is related to ν as

$$\mu = -(1 - \delta^{-1})/\nu. \tag{2.19}$$

It should be remarked here that w also satisfies the following equation:

$$(2\bar{u}_x - 1)w_t + (2\bar{u}_t - 1)w_x + T w_{tx} + 2w_t w_x = 0. \tag{2.20}$$

To show this, we first introduce (2.9) into (1.1) and then integrate once with respect to x to obtain

$$\bar{u}_t + \bar{u}_x - T\bar{u}_{tx} - 2\bar{u}_t\bar{u}_x = 0, \tag{2.21a}$$

$$\bar{v}_t + \bar{v}_x - T\bar{v}_{tx} - 2\bar{v}_t\bar{v}_x = 0. \tag{2.21b}$$

Subtracting (2.21a) from (2.21b) and noting (2.12) leads to Eq. (2.20). It can be verified by direct calculation with the aid of (B10) that Eqs. (2.14), (2.16), (2.17), and (2.21) are all compatible.

Remark 1: One can use the BT presented here to generate multisoliton solutions for Eq. (1.1). However, since this problem has already been solved by a more direct method on the basis of Eq. (2.5),[1] we restrict ourselves only on the one-soliton solution. For this case, we start with a vacuum solution of Eq. (2.3), namely, $f_+ = f_- = 1$. Then, Eqs. (2.7) reduce to the following system of linear differential equations for $g_\pm$:

$$g_{+,t} - g_{-,t} = -i(g_+ + g_-) + \nu(g_+ - g_-), \tag{2.22a}$$

$$g_{+,x} + g_{-,x} = -i(1-\delta^{-1})(g_+ - g_-) + \mu(g_+ + g_-), \tag{2.22b}$$

$$g_{+,tx} = \mu g_{+,t} + \nu g_{+,x} + \kappa g_+, \tag{2.22c}$$

$$g_{-,tx} = \mu g_{-,t} + \nu g_{-,x} + \kappa g_-, \tag{2.22d}$$

It is easy to see that Eqs. (2.22) exhibit the solutions of the forms

$$g_+ = \alpha \cosh(\gamma/2\delta)(x - ct - x_{01} - i\delta), \tag{2.23a}$$

$$g_- = \beta \cosh(\gamma/2\delta)(x - ct - x_{01} + i\delta), \tag{2.23b}$$

with the choice of the parameters

$$\kappa = -(\mu\nu + 1 - \delta^{-1}), \tag{2.23c}$$

$$\mu = i\left[\left(1 - \delta^{-1} + \frac{\gamma}{2\delta}\cot\frac{\gamma}{2}\right)\times\left(1 - \delta^{-1} - \frac{\gamma}{2\delta}\tan\frac{\gamma}{2}\right)\right]^{1/2}, \tag{2.23d}$$

$$\nu = i\left[\left(1 - \frac{c\gamma}{2\delta}\cot\frac{\gamma}{2}\right)\left(1 + \frac{c\gamma}{2\delta}\tan\frac{\gamma}{2}\right)\right]^{1/2}. \tag{2.23e}$$

Here, c is a propagation velocity of the soliton given by

$$c = (1 - \delta^{-1} + \delta^{-1}\gamma\cot\gamma)^{-1}, \tag{2.23f}$$

α and β are constants with the ratio

$$\frac{\alpha}{\beta} = -\left(1 - \delta^{-1} - i\mu + \frac{\gamma}{2\delta}\cot\frac{\gamma}{2}\right)\times\left(1 - \delta^{-1} + i\mu + \frac{\gamma}{2\delta}\cot\frac{\gamma}{2}\right)^{-1}, \tag{2.23g}$$

and γ is an arbitrary constant within the range $0 < \gamma < \pi$. It then follows from (2.1a) and (2.23) that

$$u_1 = \frac{(\gamma/2\delta)\sin\gamma}{\cosh(\gamma/\delta)(x - ct - x_{01}) + \cos\gamma}, \tag{2.24}$$

which is nothing but a one-soliton solution of Eq. (1.1).[1]

B. Conservation laws

In this subsection, we derive an infinite number of conservation laws of Eq. (1.1). For the purpose, it is most straightforward to employ the results obtained in subsection B 2. We first put

$$w = w_0 + \sum_{j=1}^{\infty} w_j \epsilon^j \quad (w_j(\pm\infty,t) = 0,\ j = 1,2,\ldots), \tag{2.25a}$$

$$\epsilon = -2\nu/(1 - \delta^{-1}). \tag{2.25b}$$

Then, Eq. (2.15) implies that

$$I_j \equiv \int_{-\infty}^{\infty} w_j\, dx \quad (j = 1,2,\ldots), \tag{2.26}$$

are conserved quantities. Substitution of (2.19) and (2.25b) into (2.16) yields

$$2i(w-w_0)=\ln\{1+(\alpha\epsilon)^2/4+i\epsilon(1+i\alpha\epsilon/2)\times[u+(i/2\delta)(w-w_0)+\hat{P}_-w_x]\}-\ln\{1+(\alpha\epsilon)^2/4-i\epsilon(1-i\alpha\epsilon/2)\times[u-(i/2\delta)(w-w_0)+\hat{P}_+w_x]\},\tag{2.27}$$

where $\alpha\equiv 1-\delta^{-1}$. Finally, substituting (2.25a) into (2.27) and comparing the coefficients of ϵ^j on both sides of (2.27), we can derive w_j successively by means of purely algebraic procedure. It is easily seen that only I_{2j+1} ($j=0, 1, \ldots$) survive and I_{2j} ($j=1, 2, \ldots$) vanish identically. Indeed, the first three of I_{2j+1} read in the forms:

$$I_1=\int_{-\infty}^{\infty}u\,dx,\tag{2.28a}$$

$$I_3=-\int_{-\infty}^{\infty}\{\tfrac{1}{3}u^3-\tfrac{1}{2}u^2+\tfrac{1}{2}uTu_x+\tfrac{1}{4}(1-\delta^{-1})u\}dx,\tag{2.28b}$$

$$I_5=\int_{-\infty}^{\infty}\{\tfrac{1}{5}u^5-\tfrac{1}{6}(3+\delta^{-1})u^4+\tfrac{1}{8}[\tfrac{16}{3}u^3-8u^2+(4-3\delta^{-1}-\delta^{-2})u-u_{xx}]Tu_x+\tfrac{1}{12}(6+\delta^{-1}-\delta^{-2})u^3+\tfrac{1}{4}(2u-1+\delta^{-1})(Tu_x)^2-\tfrac{1}{8}(2-\delta^{-1}-\delta^{-2})u^2+\tfrac{1}{8}(2u-1)u_x^2+\tfrac{1}{16}(1-\delta^{-1})^2u\}dx.\tag{2.28c}$$

It is observed that I_{2j+1} always includes a term u^{2j+1}. The lack of I_{2j} in (2.28) is a remarkable feature of conservation laws when compared with those of Eq. (1.5). In fact, Eq. (1.5) has a conserved density that includes a term u^j in I_j for all j (Ref. 8). Finally, it is worthwhile to remark that Eq. (1.1) also has an independent conserved quantity of the form

$$J_1=\int_{-\infty}^{\infty}dx\int_x^{\infty}u_t\,dy,\tag{2.29}$$

in addition to (2.28). The constancy of (2.29) in time is verified by using Eq. (2.35) below and (B5).

Remark 2: The evaluation of the jth conservation law for pure N-soliton solution of Eq. (1.1) is an interesting problem since this may provide an approximate method to obtain amplitudes of solitons evolving from arbitrary initial conditions. However, we have not as yet suitable procedure for the purpose. The main difficulty is found to arise due to the right-hand side of the formula (B7). Indeed, if this term vanishes, integration of Eq. (2.14) twice with respect to x would yield a desired result. In order to overcome the difficulty, we must probably rely on the IST method. Nevertheless, for the special cases for both deep- and shallow-water limits, the evaluation of I_j can be performed completely as will be demonstrated in Sec. III B 2 and Sec. IV B 2, respectively.

C. Inverse scattering transform

The IST of Eq. (1.1) is easily derived by employing a standard procedure in the bilinear formalism.[11,12] First, we define the wave functions $\psi_\pm$ by the relations

$$\psi_+=g_+/f_+,\tag{2.30a}$$

$$\psi_-=g_-/f_-.\tag{2.30b}$$

Substituting (2.30) into (2.7) and using (2.9) and (2.10), we find the following system of linear differential equations:

$$(\psi_++\psi_-)_x=i(2\bar{u}_x-1+\delta^{-1})(\psi_+-\psi_-)+\mu(\psi_++\psi_-),\tag{2.31a}$$

$$(\psi_+-\psi_-)_t=i(2\bar{u}_t-1)(\psi_++\psi_-)+\nu(\psi_+-\psi_-),\tag{2.31b}$$

$$\psi_{+,tx}=\mu\psi_{+,t}+\nu\psi_{+,x}-(2\delta^{-1}\bar{u}_t-4i\hat{P}_-\bar{u}_{tx}-\kappa)\psi_+,\tag{2.31c}$$

$$\psi_{-,tx}=\mu\psi_{-,t}+\nu\psi_{-,x}-(2\delta^{-1}\bar{u}_t+4i\hat{P}_+\bar{u}_{tx}-\kappa)\psi_-.\tag{2.31d}$$

It is easily verified by a cross differentiation, namely, $\psi_{+,tx}=\psi_{+,xt}$ that Eqs. (2.31) yield Eq. (1.1) as a compatibility condition. Notice also that Eqs. (2.31) must be complimented by the relation

$$(1+iT)\ln\frac{\psi_+}{\psi_{+,0}}-i\delta^{-1}\int_{-\infty}^{x}\ln\frac{\psi_+}{\psi_{+,0}}dx=(-1+iT)\ln\frac{\psi_-}{\psi_{-,0}}-i\delta^{-1}\int_{-\infty}^{x}\ln\frac{\psi_-}{\psi_{-,0}}dx,\tag{2.32}$$

which stems from (2.1), (2.2), and (2.30), where

$$\psi_{+,0}=g_{+,0}/f_{+,0}=g_+(-\infty,t)/f_+(-\infty,t),\tag{2.33a}$$

$$\psi_{-,0}=g_{-,0}/f_{-,0}=g_-(-\infty,t)/f_-(-\infty,t).\tag{2.33b}$$

A system of equations (2.31) and (2.32) constitute a complete set of the IST for Eq. (1.1).

Remark 3: One final comment to be noted here is concerned with the initial value of u_t. It can be obtained from $u(x,0)$ as follows. Let us introduce a function v through the relation

$$v=\int_x^{\infty}u_t\,dx,\tag{2.34}$$

and introduce (2.34) into Eq. (1.1). Integrating the resultant equation, one finds

$$-v+u+2uv+Tv_x=0.\tag{2.35}$$

If we solve Eq. (2.35) for a given initial condition $u(x,0)$, we can get $v(x,0)$, from which $u_t(x,0)$ follows immediately by differentiating (2.34).

D. Painlevé test

The Painlevé test provides a useful information about integrability of given system of equations. Ablowitz, Ramani, and Segur[15] conjectured that every ordinary differential equation (ODE) obtained by an exact reduction of a partial differential equation (PDE) solvable by the IST method possesses the Painlevé property, namely, solutions for ODEs have only poles as movable singularities. Thus the Painlevé property is seen to be closely related to integrability of PDEs. A drawback of the above-mentioned procedure is that one must always reduce the PDE to the ODE. In order to overcome this point, Weiss, Tabor, and Carnevale[16] have proposed a direct method that is applicable to PDEs themselves and showed that almost all soliton equations have the Painlevé property.[16,17]

Recently, Grammaticos, Dorizzi, and Ramani[19] showed that the conjecture by Ablowitz, Ramani, and Segur can also be applied to nonlinear integrodifferential evolution equations such as the BO equation and Eq. (1.5). In this subsection, following the idea due to Grammaticos, Dorizzi, and Ramani,[19] we show that Eq. (1.1) has the Painlevé property.

In the beginning, it should be observed that the problem under consideration is essentially a two-space dimensional problem. Indeed, Eq. (1.1) may be interpreted as an equation that describes internal waves propagating in the x direction in two-layered fluids, the depth of the bottom layer being δ while that of the upper layer being very shallow compared with the former one. Under this situation, the following Laplace equation for the velocity potential V must be satisfied in fluids:

$$V_{xx}+V_{yy}=0,$$
$$V=V(x,y,t), \quad (-\infty<x<\infty, \quad -\delta\leqslant y\leqslant 0), \tag{2.36a}$$

together with the boundary condition

$$V=V_0(x,t), \quad \text{at } y=0. \tag{2.36b}$$

Although Grammaticos *et al.*[19] imposed another boundary condition $V_y=0$ at $y=-\delta$, we found it inappropriate for Eq. (1.1). Indeed, this condition is shown to be incompatible with condition (2.38) given below. Now, Eq. (1.1) is equivalent to the following system of first-order PDEs:

$$U_t+V_x=0, \quad \text{at } y=0, \tag{2.37a}$$
$$U-V-V_y+2UV=0, \quad \text{at } y=0, \tag{2.37b}$$

with a subsidiary condition

$$V_y=-TV_x, \quad \text{at } y=0, \tag{2.38}$$

where $U=U(x,y,t)$. Indeed, if we put

$$U(x,0,t)=u(x,t), \tag{2.39a}$$
$$V(x,0,t)=v(x,t), \tag{2.39b}$$

then, the above fact readily follows from (2.34) and (2.35). Before performing the Painlevé test, it should be remarked that a solution of Eq. (2.36) satisfying the boundary conditions (2.36b) and (2.38) exists and it reads explicitly in the form

$$V(x,y,t)=\frac{1}{2\pi}\int_{-\infty}^{\infty} dz\, V_0(z,t) \times\int_{-\infty}^{\infty} dk\,[\cosh ky+c(k)\sinh ky]e^{ik(x-z)}, \tag{2.40a}$$

where

$$c(k)=\coth k\delta-(k\delta)^{-1}. \tag{2.40b}$$

One can verify (2.40) by using the Fourier transform method and the formula (B4).

Now, let us begin the Painlevé test. First, note that the general solution of Eq. (2.36a) is readily written as

$$V=F(x+iy,t)+G(x-iy,t), \tag{2.41}$$

where F and G are arbitrary functions. The above expression implies that the singularities of Eq. (2.36a) propagate along the characteristics $x-iy=\chi(t)$ and $x+iy=\phi(t)$. Next, consider the singularity manifold $\psi=x+iy-\phi(t)$ where $\phi(t)$ is an arbitrary function of t and expand U and V around the singularity $\psi=0$ as

$$U=\sum_{j=0}^{\infty} U_j\psi^{j-m} \quad (U_j=U_j(t), j=0,1,\ldots), \tag{2.42a}$$

$$V=\sum_{j=0}^{\infty} V_j\psi^{j-n}+G(x-iy,t) \quad (V_j=V_j(t), j=0,1,\ldots), \tag{2.42b}$$

where m and n are positive integers. These are determined by an analysis of the leading-order singularities in Eq. (2.37). The result is

$$m=n=1 \tag{2.43}$$

Finally, substituting (2.42) with (2.43) into (2.37), expanding G around $\psi=0$ as

$$G=\sum_{j=0}^{\infty}\frac{G^{(j)}}{j!}\psi^j \quad \left(G^{(j)}\equiv\left.\frac{\partial^j G(x,t)}{\partial x^j}\right|_{x=\phi-iy}\right), \tag{2.44}$$

and then taking the limit $y\to 0$, we can determine the coefficients U_j and V_j successively by balancing various powers in ψ^j. We quote only the final results as follows:

$$U_0=-i/2, \tag{2.45a}$$
$$V_0=-i\phi_t/2, \tag{2.45b}$$
$$V_1=-\phi_t U_1-G^{(0)}+\phi_t/2-\tfrac{1}{2}, \tag{2.45c}$$
$$(j+1)\phi_t U_{j+2}-(j+1)V_{j+2}=U_{j+1,t}+G^{(j+1)}/j! \quad (j\geqslant 0), \tag{2.45d}$$
$$\phi_t U_{j+2}+(j+2)V_{j+2}=-i(U_{j+1}-V_{j+1})+iG^{(j)}/j!-jG^{(j+1)}/(j+1)! -2i\sum_{k=0}^{j} U_{j-k+1}(V_{k+1}+G^{(k)}/k!) \quad (j\geqslant 0). \tag{2.45e}$$

In these expressions, U_1 is taken to be an arbitrary function, which means that a resonance condition for U_1 is satisfied automatically. From (2.45d) and (2.45e), we see that U_j and V_j $(j\geqslant 2)$ are uniquely expressed in terms of the two arbitrary functions ϕ and U_1. Hence, the expansions (2.42) are of Painlevé type. This completes the proof that Eq. (1.1) has a Painlevé property.

Before concluding this subsection, we briefly discuss both the deep- and shallow-water limits of the results presented here. For the deep-water limit $\delta \to \infty$, it follows from (2.40b) that

$$c(k) = \mathrm{sgn}(k) + O(\delta^{-1}). \tag{2.46}$$

Substituting (2.46) into (2.40a), we obtain

$$\begin{aligned} V_y|_{y=0} &= \frac{1}{2\pi}\int_{-\infty}^{\infty} dz\, V_0(z,t)\int_{-\infty}^{\infty} dk\,|k|e^{ik(x-z)} \\ &= -HV_{0,x}(x,t) = -HV_x|_{y=0}, \end{aligned} \tag{2.47}$$

where the formula (B11) has been used. Then, Eq. (1.2) follows by combining Eqs. (2.37) and (2.47).

For the shallow-water limit $\delta \to 0$, since

$$c(k) = k\delta/3 + O(\delta^3), \tag{2.48}$$

the relation corresponding to (2.47) becomes

$$\begin{aligned} V_y|_{y=0} &= \frac{\delta}{6\pi}\int_{-\infty}^{\infty} dz\, V_0(z,t)\int_{-\infty}^{\infty} dk\, k^2 e^{ik(x-z)} \\ &= -(\delta/3)V_{xx}|_{y=0}. \end{aligned} \tag{2.49}$$

Thus we find that Eqs. (2.37) and (2.49) together with the scalings (1.3) reproduce Eq. (1.4). The above two limits imply that Eqs. (1.2) and (1.4) also possess the Painlevé property.

III. STUDY OF EQ. (1.2): DEEP-WATER LIMIT

In this section, we consider the deep-water limit $\delta \to \infty$ of various results presented in Sec. II. Since the limiting procedure can be performed quite simply, we shall be concerned only with conservation laws and related topics.

A. Conservation laws

In order to obtain conservation laws, we use the BT. The time part and the space part of the BT are derived from (2.14) and (2.16), respectively, by taking the limit $\delta \to \infty$. The results are expressed as follows:

$$w_{tx} - w_t Hw_x - w_x Hw_t = \mu w_t + \nu w_x - 2\bar{u}_{tx}, \tag{3.1}$$

$$e^{2iw} = -\frac{2iP_- w_x + 2iu - i + \mu}{2iP_+ w_x + 2iu - i - \mu}, \tag{3.2a}$$

where $P_\pm$ are projection operators given by

$$P_\pm = \tfrac{1}{2}(1 \pm iH). \tag{3.2b}$$

If we impose the boundary conditions

$$w(+\infty,t) = w(-\infty,t) = w_0, \tag{3.3}$$

(2.18) and (2.19) reduce, respectively, to

$$w_0 = (1/2i)\ln[(\nu - i)/(\nu + i)], \tag{3.4}$$

$$\mu = -1/\nu. \tag{3.5}$$

It is obvious from (3.1), (3.3), and (B12) that the quantity

$$\int_{-\infty}^{\infty} (w - w_0)\,dx$$

is conserved.

Let us now derive an infinite number of conservation laws of Eq. (1.2). We first expand w as

$$w = w_0 + \sum_{j=1}^{\infty} w_j \epsilon^j, \tag{3.6a}$$

where

$$\epsilon = -2\nu, \tag{3.6b}$$

and substitute (3.6) into (3.2). It turns out by using (3.4), (3.5), and (3.6b) that

$$\begin{aligned} 2i(w - w_0) = \ln[1 + \epsilon^2/4 + i\epsilon(1 + i\epsilon/2)(u + P_- w_x)] \\ - \ln[1 + \epsilon^2/4 - i\epsilon(1 - i\epsilon/2) \\ \times (u + P_+ w_x)]. \end{aligned} \tag{3.7}$$

One can also derive (3.7) directly from (2.27) by noting $\hat{P}_\pm \to P_\pm$ and $\alpha \to 1$ in the limit of $\delta \to \infty$. Now, $w_j\,(j = 1,2,\ldots)$ are determined successively by comparing the same powers of ϵ^j on both sides of (3.7). The first three of the nontrivial conservation laws are written explicitly as follows:

$$I_1 = \int_{-\infty}^{\infty} u\,dx, \tag{3.8a}$$

$$I_3 = -\int_{-\infty}^{\infty} \left[\tfrac{1}{3}(u - \tfrac{1}{2})^3 + \tfrac{1}{2}uHu_x + \tfrac{1}{24}\right]dx, \tag{3.8b}$$

$$\begin{aligned} I_5 = \int_{-\infty}^{\infty} \big[\tfrac{1}{5}(u - \tfrac{1}{2})^5 + \tfrac{1}{8}(\tfrac{16}{3}u^3 - 8u^2 + 4u - u_{xx})Hu_x \\ + \tfrac{1}{2}u(Hu_x)^2 + \tfrac{1}{8}(2u - 3)u_x^2 + \tfrac{1}{160}\big]dx. \end{aligned} \tag{3.8c}$$

The above expressions are also obtained directly by taking the limit $\delta \to \infty$ in (2.28).

B. Evaluation of conservation laws

In this subsection, we evaluate conservation laws for N-soliton solution of Eq. (1.2). A new method developed here is independent of the IST, and is based on a previous work by the author.[20] First, we derive the corresponding result for the one-soliton solution, namely,[1]

$$u = u_1 = a/[a^2(x - ct - x_{01})^2 + 1], \tag{3.9a}$$

where a is an amplitude such as $0 < a < 1$, x_{01} is a phase constant and c is a propagation velocity given by

$$c = (1 - a)^{-1}. \tag{3.9b}$$

For the above one-soliton solution, it is obvious that w must have the following simple functional dependence on t and x

$$w = w(x - ct - x_{01}), \tag{3.10}$$

so that

$$w_t = -cw_x. \tag{3.11}$$

Substitution of (3.11) into (3.1) leads, after integrating once with respect to x, to

$$w_x - 2\int_{-\infty}^{x} w_y Hw_y\,dy = \left(\mu - \frac{\nu}{c}\right)(w - w_0) - 2u_1. \tag{3.12}$$

Integrating once again and using (3.5), (3.6b), and (B16), one arrives at the formula

$$\int_{-\infty}^{\infty} (w - w_0)\,dx = \frac{\pi\epsilon}{1 - (a - 1)(\epsilon/4)^2}. \tag{3.13}$$

This expression is interpreted as a generating function for conservation laws. The final step for obtaining I_j consists of substitution of (3.6) into (3.13) to yield

$$\sum_{j=1}^{\infty} I_j \epsilon^j = \pi \sum_{j=1}^{\infty} 2^{-2(j-1)}(a-1)^{j-1}\epsilon^{2j-1}, \quad (3.14)$$

and comparison of the coefficients of ϵ^j on both sides of (3.14). The result is very simple and it reads as follows:

$$I_{2j+1} = \pi(a-1)^j/2^{2j} \quad (j=0,1,\ldots), \quad (3.15a)$$

$$I_{2j} = 0 \quad (j=1,2,\ldots). \quad (3.15b)$$

The generalization of (3.15) to N-soliton solution u_N is done quite easily since after lapse of large time, u_N is represented by a superposition of one-soliton solutions:[1]

$$u_N \sim \sum_{n=1}^{N} \frac{a_n}{a_n^2(x-c_n t - x_{0n})^2 + 1}, \quad (3.16a)$$

with

$$c_n = (1-a_n)^{-1}, \quad 0<a_n<1 \quad (n=1,2,\ldots). \quad (3.16b)$$

Thus the relations corresponding to (3.15) become

$$I_{2j+1} = \pi \sum_{n=1}^{N} \frac{(a_n-1)^j}{2^{2j}} \quad (j=0,1,\ldots), \quad (3.17a)$$

$$I_{2j} = 0 \quad (j=1,2,\ldots). \quad (3.17b)$$

C. Initial condition evolving into pure solitons

In order to solve the initial value problem of Eq. (1.2) for arbitrary initial conditions, one must solve the IST equations. However, if we use (3.17) and an explicit form of N-soliton solution, we can find an initial condition evolving into pure solitons after lapse of large time. For the BO equation, a useful method has already been developed for the purpose.[21] In the present case, it has been conjectured that the initial condition of the form

$$u(x,0) = N\lambda/(x^2+\lambda^2), \quad (3.18)$$

would evolve into pure N solitons, where λ is a positive constant.[13] We now proceed to verify the conjecture at least up to $N=4$. Let an N-soliton solution of Eq. (1.2) expressed in terms of bilinear variable be $f_N(x,t)$. Then, it is obvious that the initial condition (3.18) evolves into pure N solitons if the equation

$$f_N(x,0) = (x-i\lambda)^N, \quad (3.19)$$

holds identically for arbitrary values of x. For the first few N, the explicit forms of $f_N(x,0)$ are given as follows:[1]

$$f_2 = \xi_1\xi_2 - B_{12}, \quad (3.20a)$$

$$f_3 = \xi_1\xi_2\xi_3 - B_{23}\xi_1 - B_{13}\xi_2 - B_{12}\xi_3, \quad (3.20b)$$

$$f_4 = \xi_1\xi_2\xi_3\xi_4 - B_{34}\xi_1\xi_2 - B_{24}\xi_1\xi_3 - B_{23}\xi_1\xi_4 - B_{14}\xi_2\xi_3 - B_{13}\xi_2\xi_4 - B_{12}\xi_3\xi_4 + B_{12}B_{34} + B_{13}B_{24} + B_{14}B_{23}, \quad (3.20c)$$

where

$$\xi_j = x - ia_j^{-1} \quad (j=1,2,\ldots,N), \quad (3.20d)$$

$$B_{jk} = 2(2-a_j-a_k)/(a_j-a_k)^2 \quad (j\neq k, j,k=1,2,\ldots,N), \quad (3.20e)$$

$$0<a_j<1 \quad (j=1,2,\ldots,N). \quad (3.20f)$$

In general, it is shown that f_N can be expressed in terms of Pfaffians.[14] One can observe by comparing the coefficients of x^j $(j=0,1,\ldots,N-1)$ on both sides of Eq. (3.19) that it yields a system of N algebraic equations for N unknowns a_1, a_2, ..., a_N and hence there exist solutions.

A general method to obtain a_j is as follows. First, introduce the following fundamental symmetric functions of a_j $(j=1,2,\ldots,N)$:

$$s_1 = \sum_{j=1}^{N} a_j, \quad (3.21a)$$

$$s_2 = \sum_{\substack{j,k=1 \\ (j<k)}}^{N} a_j a_k, \quad (3.21b)$$

$$\vdots$$

$$s_N = \prod_{j=1}^{N} a_j, \quad (3.21c)$$

and then rewrite the algebraic equations in terms of these new variables. This is always possible because of the symmetry property of N-soliton solution with respect to a_j $(j=1,2,\ldots,N)$. By solving these equations, s_j $(j=1,2,\ldots,N)$ are expressed as functions of λ. Finally, a_j $(j=1,2,\ldots,N)$ are determined from the algebraic equation of order N

$$\sum_{j=0}^{N} (-1)^j s_j A^{N-j} = 0 \quad (s_0=1). \quad (3.22)$$

It should be remarked, however, that certain restriction must be imposed on λ to satisfy conditions (3.20f). It is also interesting to observe that if (3.18) evolves into pure N solitons, s_j $(j=1,2,\ldots,N)$ are calculated directly from the relations

$$\pi \sum_{n=1}^{N} (a_n-1)^j/2^{2j} = I_{2j+1}(0) \quad (j=1,2,\ldots), \quad (3.23)$$

which stem from (3.17). Here, $I_{2j+1}(0)$ are conserved quantities evaluated at an initial time $t=0$ by using (3.18). One may also employ (3.23) to obtain approximate values of a_n for general initial conditions which would evolve into a train of solitons and tail parts.

Now, we write down the results for $N=2,3,4$.

(a) $N=2$

In this case, (3.22) becomes

$$A^2 - 2\lambda^{-2}(2\lambda-1)A + \lambda^{-3}(2\lambda-1) = 0, \quad (3.24)$$

and a_1 and a_2 are given explicitly by

$$a_1 = \{2\lambda - 1 - [(\lambda-1)(2\lambda-1)]^{1/2}\}/\lambda^2, \quad (3.25a)$$

$$a_2 = \{2\lambda - 1 + [(\lambda-1)(2\lambda-1)]^{1/2}\}/\lambda^2. \quad (3.25b)$$

The condition for λ is now expressed as

$$\lambda > \lambda_m^{(2)} \cong 2.62, \quad (3.26)$$

where $\lambda_m^{(2)}$ is the largest root of the algebraic equation

$$\lambda^2 - 3\lambda + 1 = 0. \quad (3.27)$$

(b) $N=3$

$$A^3 - 9\lambda^{-2}(\lambda-1)A^2 + \tfrac{9}{4}\lambda^{-4}(8\lambda^2 - 12\lambda + 3)A - \tfrac{3}{4}\lambda^{-5}(8\lambda^2 - 12\lambda + 3) = 0, \quad (3.28)$$

$$\lambda > \lambda_m^{(3)} \cong 4.36. \tag{3.29}$$

Here, $\lambda_m^{(3)}$ is the largest root of the algebraic equation

$$\lambda^3 - 6\lambda^2 + \tfrac{15}{2}\lambda - \tfrac{3}{2} = 0. \tag{3.30}$$

(c) $N=4$

$$A^4 - 8\lambda^{-2}(2\lambda - 3)A^3 + 8\lambda^{-4}(4\lambda^2 - 10\lambda + 5)A^2 - 12\lambda^{-6}(8\lambda^3 - 24\lambda^2 + 18\lambda - 3)A + 3\lambda^{-7}(8\lambda^3 - 24\lambda^2 + 18\lambda - 3) = 0, \tag{3.31}$$

$$\lambda > \lambda_m^{(4)} \cong 6.15. \tag{3.32}$$

Here, $\lambda_m^{(4)}$ is the largest root of the algebraic equation

$$\lambda^4 - 10\lambda^3 + 27\lambda^2 - 21\lambda + 3 = 0. \tag{3.33}$$

For $N=3,4$, the amplitudes may be written explicitly by using the famous formulas of Cardano and Ferrari. Instead of doing this, we have presented numerical values of a_j for some λ in Table I($N=3$) and Table II($N=4$), where the amplitudes are ordered such that $a_j < a_k$ for $j<k$.

Although the above procedure becomes complicated for large N, asymptotic expressions of the amplitudes a_j ($j=1,2,\ldots,N$) for large values of λ are derived easily. To show this, by observing (3.25) we put

$$x = i\lambda y, \tag{3.34a}$$

$$a_j = \tilde{a}_j/\lambda \quad (j=1,2,\ldots,N), \tag{3.34b}$$

in (3.19) and take the limit $\lambda \to \infty$. Then, (3.19) becomes

$$\tilde{f}_N(y,0) = (y-i)^N, \tag{3.35}$$

where $\tilde{f}_N$ is an N-soliton solution of the BO equation[12,22] given by

$$\tilde{f}_N(y,0) = \det M, \tag{3.36a}$$

with the $N \times N$ matrix M:

$$M = (m_{jk}) = \begin{cases} y - i\tilde{a}_j^{-1} & (j=k), \\ 2/(\tilde{a}_j - \tilde{a}_k) & (j \neq k). \end{cases} \tag{3.36b}$$

Thus, we see from Ref. 21 that $\tilde{a}_j$ ($j=1,2,\ldots,N$) coincide with the N roots of the Laguerre polynomial of order N:

$$L_N(\tilde{A}) \equiv \sum_{j=0}^{N} (-1)^j \binom{N}{j} \frac{\tilde{A}^j}{j!} = 0, \tag{3.37a}$$

where $\binom{N}{j}$ is the binomial coefficient

$$\binom{N}{j} = \frac{N!}{(N-j)!\,j!}. \tag{3.37b}$$

For the special case of $N=2$, the above results are easily confirmed with the aid of (3.25), (3.34b), and (3.37).

One final remark to be noted here is that the equation corresponding to Eq. (2.35), namely,

$$-v + u + 2uv + Hv_x = 0, \tag{3.38}$$

can be solved explicitly for the initial condition (3.18). We quote only the results. The solution of Eq. (3.38) is expressed in the form

$$v(x,0) = \sum_{j=1}^{N} \frac{\alpha_j}{(x^2+\lambda^2)^j}. \tag{3.39}$$

Here, α_j ($j=1,2,\ldots,N$) are determined uniquely from the following system of linear algebraic equations

$$(1-\lambda^{-1})\alpha_1 - \sum_{r=1}^{N-1} \frac{(2r)!}{(r!)^2 2^{2r} \lambda^{2r+1}} \alpha_{r+1} = N\lambda, \tag{3.40a}$$

$$2\lambda(N-j)\alpha_j - [1-(j+1)/\lambda]\alpha_{j+1} + \sum_{r=j+2}^{N} \frac{[2(r-j-1)]!(r-j)r}{[(r-j)!]^2 2^{2(r-j-1)} \lambda^{2(r-j)-1}} \times \alpha_r = 0 \quad (j=1,2,\ldots,N-1). \tag{3.40b}$$

If we write these equations in a matrix form as

$$\hat{A}\alpha = \beta, \tag{3.41}$$

with the column vectors $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_N)^t$ and $\beta = (N\lambda, 0, \ldots, 0)^t$ and an $N \times N$ matrix $\hat{A}$, then the equation $\det \hat{A} = 0$ yields an algebraic equation of order N for λ. Let the largest root of the equation be $\lambda_m^{(N)}$. It is now conjectured that the restriction for λ would be $\lambda > \lambda_m^{(N)}$. The conjecture is found to be true for $N=2,3,4$ for which the explicit forms of the algebraic equations are already given, respectively, by (3.27), (3.30), and (3.33). It is an interesting open problem to verify the conjecture for general N.

IV. STUDY OF EQ. (1.4): SHALLOW-WATER LIMIT

In this section, the shallow-water limit $\delta \to 0$ of the results presented in Sec. II is considered. Although the limiting procedure is somewhat complicated in comparison with the deep-water limit, it can be done straightforwardly. Hence, we shall not enter into the details of the derivations of the BT and the IST and describe only the final results. However, the properties of conservation laws are investigated in detail.

A. Conservation laws

We first derive the BT as a first step to obtain conservation laws. In the limit $\delta \to 0$, the function defined by (2.12) is expressed with the aid of (1.3), (2.1b), (2.1c), and (2.9) as

TABLE I. Amplitudes of solitons for $N=3$.

λ	a_1	a_2	a_3
5.0	0.086	0.432	0.922
7.0	0.061	0.313	0.728
9.0	0.047	0.246	0.596
11.0	0.038	0.203	0.503
13.0	0.032	0.172	0.434

TABLE II. Amplitudes of solitons for $N=4$.

λ	a_1	a_2	a_3	a_4
7.0	0.047	0.244	0.568	0.936
9.0	0.037	0.190	0.455	0.799
11.0	0.030	0.156	0.380	0.690
13.0	0.025	0.133	0.325	0.606
15.0	0.022	0.115	0.285	0.538

$$w = (3\delta)^{1/2}\frac{\partial}{\partial\xi}\ln\frac{g}{f} + O(\delta). \tag{4.1}$$

Suggested by this expression, we introduce the new variable W through the relation

$$W = \left(\frac{\delta}{3}\right)^{-1/2} w = 3\frac{\partial}{\partial\xi}\ln\frac{g}{f}, \tag{4.2}$$

together with the scaled parameters $\tilde{\mu}$ and $\tilde{\nu}$

$$\tilde{\mu} = -(3\delta)^{-1/2}\mu, \tag{4.3a}$$

$$\tilde{\nu} = (3\delta)^{1/2}\nu. \tag{4.3b}$$

Substituting (1.3), (4.2), (4.3), and (B8) into (2.17) and expanding with δ, one finds after neglecting higher-order terms in δ:

$$W_\tau + \frac{1}{3}W\int_{-\infty}^{\infty} W_\tau\, d\xi = -2\bar{u}_\tau + \frac{1}{3}\tilde{\nu}W + 1. \tag{4.4}$$

The boundary value $W_0 \equiv W_0(\pm\infty,\tau)$ is obtained by using (2.18), (4.2), and (4.3b) as

$$W_0 = -3/\tilde{\nu}. \tag{4.5}$$

Equation (4.4) with (4.5) represents the time part of the BT. If we take $\xi \to \infty$ in (4.4) and use (4.5), we obtain

$$\int_{-\infty}^{\infty} W_\tau\, d\xi = 0, \tag{4.6}$$

and this relation implies that the quantity

$$\tilde{I} \equiv \int_{-\infty}^{\infty}(W - W_0)d\xi \tag{4.7}$$

is conserved.

The space part of the BT follows similarly by introducing (2.18), (2.19), (4.2), (4.3), (4.5), and (B5) into (2.16) and performing the limiting procedure as follows:

$$W_{\xi\xi} + (W_\xi + 2u - 1)W + \tfrac{1}{3}W^3 = 3(\tilde{\nu}^{-1} - \tilde{\nu}^{-3}). \tag{4.8}$$

It is useful to note that Eq. (4.4) and Eq. (4.8) can also be derived from a pair of the BT of Eq. (1.4) expressed in terms of bilinear variables[11] with the aid of (4.2) and the formulas (A6)–(A8).

In order to derive conservation laws, we first introduce the following quantities

$$W_j = (\delta/3)^{(j-1)/2} w_j \quad (j = 1,2,\ldots), \tag{4.9a}$$

$$\tilde{\epsilon} = (\delta/3)^{-1/2}\epsilon. \tag{4.9b}$$

It then follows from (2.25), (4.2), and (4.9) that

$$W = W_0 + \sum_{j=1}^{\infty} W_j \tilde{\epsilon}^j. \tag{4.10}$$

Notice that (2.25b), (4.3b), and (4.9b) yield a relation

$$\tilde{\epsilon} = 2\tilde{\nu}. \tag{4.11}$$

Expanding $\tilde{I}$ as

$$\tilde{I} = \sum_{j=1}^{\infty} \tilde{I}_j \tilde{\epsilon}^j, \tag{4.12}$$

and substituting (4.10) and (4.12) into (4.7), one finds that

$$\tilde{I}_j = \int_{-\infty}^{\infty} W_j\, d\xi \quad (j = 1,2,\ldots), \tag{4.13}$$

are conserved quantities. If we use (4.9) and

$$W_0 = -3/\tilde{\nu} = -6/\tilde{\epsilon}, \tag{4.14}$$

Eq. (4.8) is rewritten in the form

$$W_{\xi\xi} + (W_\xi + 2u)W + \tfrac{1}{3}(W - W_0)^3 - (2/\tilde{\epsilon})(W - W_0)^2 - \left(1 - \frac{12}{\tilde{\epsilon}^2}\right)(W - W_0) = 0. \tag{4.15}$$

Finally, substituting (4.10) into (4.15) and taking the coefficients of $\tilde{\epsilon}^j (j = 1,2,\ldots)$ zero, we obtain the recursion relations that generate W_j. One then finds that only $\tilde{I}_j$ with odd j survive. Explicitly up to $j = 11$, they read as follows:

$$\tilde{I}_1 = \int_{-\infty}^{\infty} u\, d\xi, \tag{4.16a}$$

$$\tilde{I}_3 = \frac{1}{12}\int_{-\infty}^{\infty} u\, d\xi, \tag{4.16b}$$

$$\tilde{I}_5 = -\frac{1}{108}\int_{-\infty}^{\infty}\left(u^3 - \frac{3}{2}u^2 - \frac{3}{2}u_\xi^2 - \frac{3}{4}u\right)d\xi, \tag{4.16c}$$

$$\tilde{I}_7 = -\frac{1}{648}\int_{-\infty}^{\infty}\left(u^4 + \frac{1}{2}u^3 - \frac{9}{2}uu_\xi^2 + \frac{3}{4}u_{\xi\xi}^2 - \frac{9}{4}u^2 - \frac{3}{2}u_\xi^2 - \frac{3}{8}u\right)d\xi, \tag{4.16d}$$

$$\tilde{I}_9 = -\frac{7}{7776}\int_{-\infty}^{\infty}\left(u^4 - \frac{4}{7}u^3 - \frac{9}{2}uu_\xi^2 + \frac{3}{4}u_{\xi\xi}^2 - \frac{9}{14}u^2 + \frac{3}{28}u_\xi^2 - \frac{3}{56}u\right)d\xi, \tag{4.16e}$$

$$\tilde{I}_{11} = \frac{1}{157464}\int_{-\infty}^{\infty}\left(u^6 - 3u^5 - \frac{75}{4}u^3u_\xi^2 - \frac{45}{16}u^4 + \frac{225}{8}u^2u_\xi^2 + 9u^2u_{\xi\xi}^2 - \frac{51}{16}u_\xi^4 + \frac{15}{4}u^3 + \frac{495}{32}uu_\xi^2 - 9uu_{\xi\xi}^2 - \frac{63}{32}uu_{\xi\xi\xi\xi}^2 + \frac{3}{2}u_{\xi\xi}^2 + \frac{63}{64}u_{\xi\xi\xi}^2 + \frac{45}{32}u^2 - \frac{135}{64}u_\xi^2 - \frac{171}{64}u_{\xi\xi}^2 + \frac{9}{64}u_{\xi\xi\xi\xi}^2 + \frac{9}{128}u\right)d\xi. \tag{4.16f}$$

These quantities can also be derived from (2.28) by introducing the scalings $\tilde{I}_j = (\delta/3)^{-j/2} I_j (j = 1,2,\ldots)$ in addition to (1.3). It is quite interesting to observe that in (4.16), only $\tilde{I}_1, \tilde{I}_5, \tilde{I}_7$, and $\tilde{I}_{11}$ are independent. Indeed, $\tilde{I}_3$ and $\tilde{I}_9$ are represented in terms of these quantities as

$$\tilde{I}_3 = \frac{1}{12}\tilde{I}_1, \tag{4.17a}$$

$$\tilde{I}_9 = \frac{1}{2304}\tilde{I}_1 - \frac{5}{48}\tilde{I}_5 + \frac{7}{12}\tilde{I}_7. \tag{4.17b}$$

The peculiar structure of these conserved quantities will be clarified in the following (see C).

B. Evaluation of conservation laws

We now develop a new method to evaluate conservation laws for N-soliton solution. For the purpose, we use actively the time part of the BT. In the beginning, we do this for a one-soliton solution of Eq. (1.4):

$$u = u_1 = 3a^2 \operatorname{sech}^2 a(\xi - c\tau - \xi_{01}) \quad [c = (1-4a^2)^{-1}]. \tag{4.18}$$

It then turns out that W has the following functional form

$$W = W(\xi - c\tau - \xi_{01}). \tag{4.19}$$

Substituting (4.19) into (4.4) and integrating once with respect to ξ, we obtain after some modifications

$$3W_\xi + (W - W_0)^2 + \left(-\frac{3}{\tilde{\nu}} + \frac{\tilde{\nu}}{c}\right)(W - W_0) + 6u_1 = 0. \tag{4.20}$$

This is an equation of the Riccati type. Owing to the second term on the left-hand side of Eq. (4.20), one cannot rely on the method developed in Sec. III B. Indeed, for Eq. (3.12) corresponding to Eq. (4.20), the second term on the left-hand side has been shown to vanish identically by integrating with respect to x from $-\infty$ to ∞. However, luckly in the present case, the difficulty is overcome as follows. We start by introducing new variables Y and η through the relations

$$W - W_0 = 3Y_\xi / Y, \tag{4.21a}$$

$$\eta = a(\xi - c\tau - \xi_{01}), \tag{4.21b}$$

and setting

$$\frac{-3}{\tilde{\nu}} + \frac{\tilde{\nu}}{c} = \frac{-6}{\tilde{\epsilon}} + \frac{(1-4a^2)\tilde{\epsilon}}{2} \equiv -6a\sigma. \tag{4.21c}$$

Then, Eq. (4.20) becomes

$$Y_{\eta\eta} - 2\sigma Y_\eta + 2\operatorname{sech}^2 \eta Y = 0. \tag{4.22}$$

Furthermore, if we make a change of the independent variable η as

$$\zeta = (1 - \tanh \eta)/2, \tag{4.23}$$

Eq. (4.22) is transformed into the form

$$\zeta(1-\zeta)Y_{\zeta\zeta} + (1 + \sigma - 2\zeta)Y_\zeta + 2Y = 0. \tag{4.24}$$

Equation (4.24) is a special case of the following hypergeometric differential equation of Gauss[23]

$$\zeta(1-\zeta)Y_{\zeta\zeta} + [\gamma - (\alpha + \beta + 1)\zeta]Y_\zeta - \alpha\beta Y = 0, \tag{4.25}$$

with $\alpha = 2, \beta = -1$, and $\gamma = \sigma + 1$. As is well known,[23] a solution of Eq. (4.25) regular at $\zeta = 0$ is represented by an infinite series

$$Y = F(\alpha,\beta,\gamma;\zeta) = \sum_{n=0}^{\infty} \frac{(\alpha)_n (\beta)_n}{(\gamma)_n} \frac{\zeta^n}{n!}, \tag{4.26a}$$

with

$$(\alpha)_n = \alpha(\alpha+1)\cdots(\alpha+n-1). \tag{4.26b}$$

If we note the boundary condition $\lim_{|\xi|\to\infty} Y = \text{const}$ which is a consequence of (4.21a), we find that an appropriate solution of Eq. (4.24) satisfying the boundary condition is written in a finite series of the form

$$Y = F(2, -1, \sigma + 1; \zeta) = 1 - [2/(\sigma+1)]\zeta. \tag{4.27}$$

Using the above expression, one can easily evaluate the quantities given by (4.13). Indeed, it follows from (4.7) and (4.21a) that

$$\begin{aligned}\tilde{I} &= [3 \ln Y]_{\eta=-\infty}^{\eta=+\infty} \\ &= [3 \ln Y]_{\zeta=1}^{\zeta=0} \\ &= 3 \ln[(\sigma+1)/(\sigma-1)].\end{aligned} \tag{4.28}$$

Substitution of (4.12) and (4.21c) into (4.28) leads to

$$\sum_{j=1}^{\infty} \tilde{I}_j \tilde{\epsilon}^j = 3 \ln\left[\frac{1 + a\tilde{\epsilon} - (1-4a^2)\tilde{\epsilon}^2/12}{1 - a\tilde{\epsilon} - (1-4a^2)\tilde{\epsilon}^2/12}\right]. \tag{4.29}$$

Expanding the right-hand side of (4.29) with $\tilde{\epsilon}$ and then comparing the coefficients of $\tilde{\epsilon}^j$ on both sides, one arrives at the final result as follows:

$$\tilde{I}_{2j+1} = 6 \sum_{r=0}^{j} \frac{\binom{2j-r+1}{r}}{2j-r+1} \times \frac{a^{2j-2r+1}(1-4a^2)^r}{12^r} \quad (j = 0,1,\ldots), \tag{4.30a}$$

$$\tilde{I}_{2j} = 0 \quad (j = 1,2,\ldots). \tag{4.30b}$$

Explicitly, the nonzero $\tilde{I}_j$ up to $j = 11$ read in the forms

$$\tilde{I}_1 = 6a, \tag{4.31a}$$

$$\tilde{I}_3 = \frac{a}{2}, \tag{4.31b}$$

$$\tilde{I}_5 = -\frac{5}{12}\left(a^5 - \frac{5}{4}a^3 - \frac{5}{16}a\right), \tag{4.31c}$$

$$\tilde{I}_7 = -\frac{2}{63}\left(a^7 - \frac{21}{16}a^3 - \frac{7}{64}a\right), \tag{4.31d}$$

$$\tilde{I}_9 = -\frac{1}{54}\left(a^7 - \frac{3}{4}a^5 - \frac{3}{8}a^3 - \frac{1}{64}a\right), \tag{4.31e}$$

$$\tilde{I}_{11} = \frac{2}{891}\left(a^{11} - \frac{11}{4}a^9 - \frac{11}{16}a^7 + \frac{143}{64}a^5 + \frac{55}{128}a^3 + \frac{11}{1024}a\right). \tag{4.31f}$$

The generalization of the above results to those for an N-soliton solution is obvious. One may simply replace (4.30a) by the expressions

$$\tilde{I}_{2j+1} = 6 \sum_{r=0}^{j} \frac{\binom{2j-r+1}{r}}{2j-r+1} \times \sum_{n=1}^{N} \frac{a_n^{2j-2r+1}(1-4a_n^2)^r}{12^r} \quad (j = 0,1,\ldots), \tag{4.32}$$

where $3a_n^2$ represents the amplitude of the nth soliton.

C. Structure of conservation laws

As easily observed from the explicit expressions (4.31) for $\tilde{I}_{2j+1}$, all these quantities are not independent. One can confirm by direct calculation that the relations (4.17) hold for (4.31). This fact provides a useful information about the structure of conservation laws. In this subsection, we show that $\tilde{I}_{2j+1}$ $(j \neq 3n+1, n = 0,1,\ldots)$ are only independent conserved quantities. In other words, if $2j+1$ is equal to an odd integer times 3, namely, $2j+1 = 3(2n+1)$ $(n = 0,1,\ldots)$, then $\tilde{I}_{2j+1}$ are expressed as a linear combination of independent conserved quantities. It is sufficient to prove this statement only for (4.31) since the corresponding result for an N-soliton solution readily follows from (4.32).

First, we note from (4.29) that $\tilde{I}_{2j+1}$ is also written in the form

$$\tilde{I}_{2j+1} = \frac{6\cdot 12^{-(2j+1)/2}}{2j+1}[(\sqrt{3}a+\sqrt{1-a^2})^{2j+1} - (\sqrt{3}a-\sqrt{1-a^2})^{2j+1}]. \tag{4.33}$$

If we put

$$a = \cos\theta, \tag{4.34}$$

then, (4.33) becomes

$$\tilde{I}_{2j+1} = \frac{2\cdot 3^{-(2j-1)/2}}{2j+1} \times\left[\cos^{2j+1}\left(\theta-\frac{\pi}{6}\right)+\cos^{2j+1}\left(\theta+\frac{\pi}{6}\right)\right]. \tag{4.35}$$

Using the well-known formulas

$$\cos^{2j+1}\theta = 2^{-2j}\sum_{r=0}^{j}\binom{2j+1}{r}\cos(2j-2r+1)\theta, \tag{4.36a}$$

$$\cos\theta_1+\cos\theta_2 = 2\cos\tfrac{1}{2}(\theta_1+\theta_2)\cos\tfrac{1}{2}(\theta_1-\theta_2), \tag{4.36b}$$

(4.35) is modified as

$$\tilde{I}_{2j+1} = \frac{2^{-2(j-1)}3^{-(2j-1)/2}}{2j+1} \times\sum_{r=0}^{j}\binom{2j+1}{r}\cos\frac{\pi}{6}(2j-2r+1) \times\cos(2j-2r+1)\theta. \tag{4.37}$$

This is a desired form. However, it is sometimes more convenient to express (4.37) in terms of a. To do so, we use the formula

$$\cos(2j+1)\theta = \sum_{r=0}^{j}\frac{(-1)^r(2j+1)}{2(2j+1-r)}\binom{2j+1-r}{r} \times(2\cos\theta)^{2j-2r+1}. \tag{4.38}$$

Substitution of (4.34) and (4.38) into (4.37) yields, after some manipulations, the following alternative expression for $\tilde{I}_{2j+1}$:

$$\tilde{I}_{2j+1} = \frac{12^{-(j-1)}}{2\sqrt{3}(2j+1)}\sum_{r=0}^{j}(2a)^{2r+1} \times\sum_{s=0}^{j-r}\frac{(-1)^s(2r+2s+1)}{2r+s+1} \times\binom{2(j-r-s)+1}{j-r-s}\binom{2r+s+1}{s} \times\cos\frac{\pi}{6}[2(r+s)+1]. \tag{4.39}$$

It is seen from (4.39) that $\tilde{I}_{2j+1}$ with $2j+1\neq 3(2n+1)$ includes a term a^{2j+1} as a maximum power of a, but for $2j+1=3(2n+1)$ it lacks the term and instead has a^{2j-1} as a corresponding term.

Under these preparations, let us now prove the dependence of conserved quantities mentioned at the beginning of this section. For the purpose, consider the following quantity:

$$\tilde{J}_{6n+3} \equiv \sum_{j=0}^{3n}\alpha_{n,j}\tilde{I}_{2j+1} = \sum_{j=0}^{3n}\alpha_{n,j}\sum_{r=0}^{j}\binom{2j+1}{j-r}c_jF_r, \tag{4.40a}$$

where

$$F_j = \cos\frac{\pi}{6}(2j+1)\cos(2j+1)\theta, \tag{4.40b}$$

$$c_j = \frac{2^{-2}(j-1)3^{-(2j-1)/2}}{2j+1}, \tag{4.40c}$$

and $\alpha_{n,j}$ are unknown constants determined later. One then modifies (4.40) in the form

$$\tilde{J}_{6n+3} = \sum_{j=0}^{3n}\sum_{r=0}^{3n-j}\alpha_{n,j+r}\binom{2(j+r)+1}{r}c_{j+r}F_j. \tag{4.41}$$

Furthermore, if $\alpha_{n,j}\,(j=0,1,\ldots,3n)$ satisfy the following system of linear algebraic equations:

$$\sum_{r=0}^{3n-j}\alpha_{n,j+r}\binom{2(j+r)+1}{r}c_{j+r} = \binom{6n+3}{3n+1-j}c_{3n+1} \quad (j=0,1,\ldots,3n), \tag{4.42}$$

then, (4.40) becomes

$$\tilde{J}_{6n+3} = \sum_{j=0}^{3n}\binom{6n+3}{3n+1-j}c_{3n+1}F_j. \tag{4.43}$$

However, one easily sees by noting $F_{3n+1}=0$ that $\tilde{J}_{6n+3}$ coincides with $\tilde{I}_{6n+3}$ and this means that $\tilde{I}_{6n+3}$ is expressed by a linear combination of $\tilde{I}_{2j+1}\,(j=0,1,\ldots,3n)$. The existance of $\alpha_{n,j}$ is obvious since the determinant constructed from the coefficients of $\alpha_{n,j}$ in (4.42) is

$$\prod_{r=0}^{3m}c_r$$

and hence never vanishes due to (4.40c). Thus we have finished the proof.

Remark 1: The similar structure of conservation laws exists[24-26] for the Sawada–Kotera(SK) equation:[27]

$$u_\tau + 180u^2u_\xi + 30(uu_{\xi\xi\xi}+u_\xi u_{\xi\xi}) + u_{\xi\xi\xi\xi\xi} = 0, \tag{4.44}$$

for which the IST problem has been partially solved.[26,28]

Remark 2: The discussion on the dependence of conservation laws developed here is based on an N-soliton solution. However, since the conservation laws hold for arbitrary initial conditions, the conclusion obtained in this section gives a necessary condition for the dependence. Nevertheless, the investigation of lower conservation laws [see (4.17), for instance] strongly suggests that it is also sufficient. This is an interesting problem to be pursued further.

Remark 3: The IST equations arise naturally from (2.31) by noting the relations

$$\psi_\pm = \frac{g_\pm}{f_\pm} = \exp\left(\mp i\delta\frac{\partial}{\partial x}\right)\frac{g}{f}. \tag{4.45}$$

If we define the wave function ψ by

$$\psi = g/f, \tag{4.46}$$

substitute (4.46), (1.3), and (4.3) into (2.31) and take the limit $\delta\to 0$, we obtain the following system of linear differential equations

$$3\psi_{\tau\xi} = \tilde{\nu}\psi_{\xi} - (2\bar{u}_{\tau} - 1)\psi, \tag{4.47a}$$

$$\psi_{\xi\xi\xi} + (2u - 1)\psi_{\xi} = \tilde{\mu}\psi, \tag{4.47b}$$

which constitute the IST of Eq. (1.4). It is important to remark that the space part of the IST (4.47b) is essentially the same form as that corresponding to the SK equation (4.44).[25,26]

V. CONCLUDING REMARKS

In this paper, starting from the bilinear form of a model equation for water waves in fluids of finite depth we have constructed the BT, an infinite number of conservation laws and the IST for the equation. Then, both the deep- and shallow-water limits of various results thus obtained have been investigated in detail. In particular, the structure of conservation laws has been clarified and it was found that it exhibits quite peculiar characteristics in comparison with those of usual water wave equations such as the BO and the KdV equations.

The most important open problem left in this paper would be to solve various IST equations. In this respect, it should be remarked that the space part of the IST equation for the shallow-water wave equation considered in Sec. IV has the same form as that for the SK equation. However, one must keep in mind that the IST problem for the SK equation has been solved only partially.[26,28] For instance, the explicit form of N-soliton solution has not been derived as yet within the framework of the IST formalism. Concerning this point, it may be instructive to note that Hirota recently found Pfaffian expressions of N-soliton solutions for the SK and related equations[29] and derived a new type of linear integral equation[30] that corresponds to the well-known Gel'fand–Levitan equation. The above-mentioned problems must be studied further in future works.

AKNOWLEDGMENT

The author wishes to express his sincere thanks to Professor M. Nishioka for continual encouragement.

APPENDIX A: FORMULAS OF BILINEAR OPERATORS

The following formulas are easily verified by direct calculations using the definition of bilinear operators (2.4), where f, f', g, and g' are functions of t and x and $\phi = \ln(f/g), \rho = \ln(fg)$:

$$D_x f\cdot f = 0, \tag{A1}$$

$$D_t(fg\cdot D_x f\cdot g) = D_x(fg\cdot D_t f\cdot g), \tag{A2}$$

$$f'g'D_x f\cdot g - (D_x f'\cdot g')fg = D_x g' f\cdot f' g, \tag{A3}$$

$$\begin{aligned} f'g'D_t D_x f\cdot g - (D_t D_x f'\cdot g')fg &= (D_t D_x f'\cdot f)g'g - f'f D_t D_x g'\cdot g \\ &\quad + D_t[g'f\cdot(D_x f'\cdot g)] + D_x[g'f\cdot(D_t f'\cdot g)], \end{aligned} \tag{A4}$$

$$\exp(i\delta D_x) f\cdot g = f(x + i\delta)g(x - i\delta), \tag{A5}$$

$$D_x f\cdot g/fg = \phi_x, \tag{A6}$$

$$D_t D_x f\cdot g/fg = \rho_{tx} + \phi_t\phi_x, \tag{A7}$$

$$D_x^3 f\cdot g/fg = \phi_{xxx} + 3\phi_x\rho_{xx} + \phi_x^3. \tag{A8}$$

APPENDIX B: PROPERTIES OF OPERATORS T AND H

The following formulas are obtained by employing the Fourier transform method together with the Chauchy residue theorem and the formulas

$$[\coth k_1\delta + \coth k_2\delta]\coth(k_1 + k_2)\delta = \coth k_1\delta \coth k_2\delta + 1, \tag{B1}$$

$$\lim_{\delta\to\infty} \coth k\delta = \operatorname{sgn}(k). \tag{B2}$$

In the following, $\tilde{f}(k)$ denotes the Fourier transform of $f(x)$:

$$\tilde{f}(k) = \frac{1}{2\pi}\int_{-\infty}^{\infty} f(x)e^{-ikx}\,dx, \tag{B3}$$

and the existence of $\tilde{f}(0)$ is assumed.

1. T operator

$$Te^{ikx} = i[\coth k\delta - (1/k\delta)]e^{ikx}, \tag{B4}$$

$$\int_{-\infty}^{\infty}(fTg + gTf)dx = 0, \tag{B5}$$

$$\int_{-\infty}^{\infty} Tf\,dx = 0, \tag{B6}$$

$$\int_{-\infty}^{\infty} dx\int_{-\infty}^{x} f_y Tf_y\,dy - \frac{1}{2\delta}\int_{-\infty}^{\infty} f^2\,dx = -\pi\delta\int_{-\infty}^{\infty}\tilde{f}(k)\tilde{f}(-k)(k/\sinh k\delta)^2 dk, \tag{B7}$$

$$Tf = \frac{\delta}{3}f_x + \frac{\delta^3}{45}f_{xxx} + \frac{2}{945}\delta^5 f_{xxxxx} + O(\delta^7). \tag{B8}$$

If we define the operator $\hat{T}$ by

$$\hat{T}f(x) = P\int_{-\infty}^{\infty}\frac{1}{2\delta}\coth\frac{\pi(y-x)}{2\delta}f(y)\,dy, \tag{B9}$$

then, for the functions such that $\lim_{k\to 0}\tilde{f}(k) = O(k)$ and $\lim_{k\to 0}\tilde{g}(k) = O(k)$

$$\hat{T}(f\hat{T}g + g\hat{T}f) = (\hat{T}f)(\hat{T}g) - fg. \tag{B10}$$

2. H operator

$$He^{ikx} = i\operatorname{sgn}(k)e^{ikx}, \tag{B11}$$

$$\int_{-\infty}^{\infty}(fHg + gHf)dx = 0, \tag{B12}$$

$$\int_{-\infty}^{\infty} Hf\,dx = 0, \tag{B13}$$

$$H^2 f = -f, \tag{B14}$$

$$H(fg) = H[(Hf)(Hg)] + fHg + gHf, \tag{B15}$$

$$\int_{-\infty}^{\infty} dx\int_{-\infty}^{x} f_y Hf_y\,dy = 0. \tag{B16}$$

[1] Y. Matsuno, J. Math. Phys. **29**, 49(1989); **30**, 241(1989).
[2] R. Hirota and J. Satsuma, J. Phys. Soc. Jpn. **40**, 611(1976).
[3] R. I. Joseph, J. Phys. A: Math. Gen. **10** L225(1977).
[4] T. Kubota, D. R. S. Ko, and D. Dobbs, J. Hydronau. **12**, 157(1978).
[5] T. B. Benjamin, J. Fluid Mech. **29**, 559(1967).
[6] H. Ono, J. Phys. Soc. Jpn. **39**, 1082(1975).
[7] Y. Matsuno, Phys. Lett. A **74**, 223(1979).
[8] J. Satsuma, M. J. Ablowitz, and Y. Kodama, Phys. Lett. A **73**, 283(1979).

[9] Y. Kodama, M. J. Ablowitz, and J. Satsuma, J. Math. Phys. **23**, 564(1982).
[10] R. Hirota, Phys. Rev. Lett. **27**, 1192(1971).
[11] R. Hirota, in *Solitons*, edited by R. K. Bullough and P. J. Caudrey (Springer-Verlag, Berlin, 1980), p. 157.
[12] Y. Matsuno, *Bilinear Transformation Method* (Academic, New York, 1984).
[13] Y. Matsuno, J. Phys. Soc. Jpn. **57**, 1577(1988).
[14] Y. Matsuno, J. Phys. Soc. Jpn. **58**, 1948(1989).
[15] M. J. Ablowitz, A. Ramani, and H. Segur, J. Math Phys. **21**, 715(1980); **21**, 1006(1980).
[16] J. Weiss, M. Tabor, and J. Carnevale, J. Math. Phys. **24**, 522(1983).
[17] J. Weiss, J. Math. Phys. **24**, 1405(1983); **25**, 13(1984); **25**, 2226(1984); **26**, 258(1985); **26**, 2174(1985); **27**, 1923(1986).
[18] A. Ramani, B. Grammaticos, and T. Bountis, Phys. Rep. **180**, 160(1989).
[19] B. Grammaticos, B. Dorizzi, and A. Ramani, Phys. Rev. Lett. **53**, 1(1984).
[20] Y. Matsuno, J. Phys. Soc. Jpn. **51**, 667(1982).
[21] Y. Matsuno, J. Phys. Soc. Jpn. **51**, 3375(1982).
[22] Y. Matsuno, J. Phys. A: Math. Gen. **12**, 619(1979).
[23] H. Hochstadt, *The Functions of Mathematical Physics* (Wiley, New York, 1971).
[24] P. J. Caudrey, R. K. Dodd, and J. D. Gibbon, Proc. R. Soc. London Ser. A **351**, 407(1977).
[25] J. Satsuma and D. J. Kaup, J. Phys. Soc. Jpn. **43**, 692(1977).
[26] D. J. Kaup, Stud. Appl. Math. **62**, 189(1980).
[27] K. Sawada and T. Kotera, Prog. Theor. Phys. **51**, 1355(1974).
[28] P. J. Caudrey, Phys. Lett. A **79**, 264(1980).
[29] R. Hirota, J. Phys. Soc. Jpn. **58**, 2285(1989).
[30] R. Hirota, J. Phys. Soc. Jpn. **58**, 2705(1989).

Exact plane-wave solutions of the coupled Maxwell–Klein–Gordon equations

A. Das, T. Biech, and D. Kay[a)]
Department of Mathematics and Statistics, Simon Fraser University, Burnaby, British Columbia V5A 1S6, Canada

(Received 9 May 1990; accepted for publication 1 August 1990)

Solutions of the classical Maxwell–Klein–Gordon equations are investigated for which the Klein–Gordon field is assumed to be $\psi(x) = \alpha e^{ip_\mu x^\mu}$. It is shown that for this class the exponential factor can be "gauged away" and the resulting system of equations can be reduced to a *single* (complicated) nonlinear equation. Furthermore, the electromagnetic four-potential field becomes *massive* "absorbing scalar particles." The steady-state (or stationary) subclass of the resulting system of equations is examined. It is proved that in absence of any magnetic field, the steady-state system does not have a solution. In the simple case for which four-potential components A^μ depend on *one* spatial coordinate, the equations are completely solved and explicitly analyzed.

I. INTRODUCTION

Exact (nontrivial) solutions of any system of interacting quantized fields are not known. A system of interacting classical fields is much easier to investigate. The initial-value problem of the classical Maxwell–Dirac equations has been studied[1] in this decade. Exact plane-wave solutions of the Maxwell–Dirac equations have been examined recently.[2] Exact solutions of the coupled Einstein–Maxwell–Klein–Gordon equations have been found[3] throughout the last 2 decades.

A particular plane-wave solution of the combined Einstein–Klein–Gordon–Dirac–Maxwell equations (where the curvature was due to the Takeno metric), has been given by Brill and Cohen.[4]

Procedures for exact integration of the differential equations in these papers bypass completely perturbative techniques that are plagued with "divergence difficulties."

In this paper, we seek solutions of Maxwell–Klein–Gordon (MKG) equations with Klein–Gordon field $\psi(x) = \alpha e^{ip_\mu x^\mu}$. In elementary wave mechanics such a wave function represents a momentum eigenstate. It turns out that for any complex wave field with "minimal electromagnetic interaction" the factor $e^{ip_\mu x^\mu}$ can be "absorbed" by a gauge transformation. Therefore, the wave function $\psi(x) = \alpha e^{ip_\mu x^\mu}$ is gauge equivalent (also physically equivalent) to the constant-valued wave function $\tilde{\psi}(x) = \alpha$. With such an assumption, we study MKG equations in Sec. III. It is proved that the electromagnetic four-potential $A^\mu(x)$ and the charge-current four-vector $j^\mu(x)$ are timelike everywhere and furthermore these vectors are constant multiples of each other. Moreover, the four-potential field $A^\mu(x)$ satisfy the Proca equations, and thus *photons acquire mass* as it were, absorbing scalar particles or mesons.

In Sec. IV, MKG equations with constant-valued wave functions are further simplified by the additional assumption that MKG fields are in steady state or stationary. It is proved that in the absence of magnetic fields, the steady-state equations have no solution. Furthermore, it is proved that if at a point of a spatial domain, the potential $A^0(x)$ attains a minimum, then solutions of the steady-state equations do not exist on that spatial domain.

In the last section, under the simplifying assumptions $\psi(x) = \alpha$ and the A^μ are functions of a single spatial coordinate, the MKG equations are completely integrated. The electromagnetic field tensor, symmetrized energy-momentum-stress tensors are explicitly computed for these solutions. Moreover, the total energy and momentum of the MKG fields inside a bounded spatial cylindrical domain are explicitly obtained.

II. NOTATION AND FIELD EQUATIONS

The flat space-time differentiable manifold is denoted by M. A Minkowski coordinate chart is used for M. The symbol $:=$ is used for definitions and the symbol $\equiv$ is used for identities. A space-time event, which is an element of M, is coordinatized by $x := (x^0, x^1, x^2, x^3)$, where x^0 denotes the time coordinate. A Greek index takes values in $\{0,1,2,3\}$ and a Roman index takes values in $\{1,2,3\}$. The signature of the metric is assumed to be -2 so that the metric tensor $[\eta_{\mu\nu}] = \mathrm{diag}[1,-1,-1,-1]$. Einstein's summation convention is used for both types of indices. The physical units are chosen so that $\hbar = c = 1$. The electromagnetic four-potential components are indicated by $A^\mu(x)$. The partial derivatives are denoted by ∂_μ. We shall usually assume that in a domain $D \subset \mathbb{R}^4$ (corresponding to a domain in M), the Klein–Gordon function ψ is of the differentiability class $\mathscr{C}^2(D;\mathbb{C})$ and the four-potential function A^μ are of the differentiability class $\mathscr{C}^3(D;\mathbb{R})$. In such a domain $D \subset \mathbb{R}^4$, the coupled MKG equations (which are Poincaré covariant as well as gauge invariant) are the following:

$$k(x) := \{[\partial^\mu + ieA^\mu(x)] \times [\partial_\mu + ieA_\mu(x)] + m^2\}\psi(x) = 0,$$
$$M^\mu(x) := \partial_\nu F^{\mu\nu}(x) - j^\mu(x) = 0, \tag{2.1}$$

[a)] Current address: Okanagan College, Kelowna, British Columbia V1Y 4X8, Canada.

 0022-2488/90/122917-04$03.00

$$F_{\mu\nu}(x) := \partial_\nu A_\mu(x) - \partial_\mu A_\nu(x),$$
$$j^\mu(x) := ie\{[\partial^\mu - ieA^\mu(x)]\bar{\psi}(x)\psi(x) - \bar{\psi}(x)[\partial^\mu + ieA^\mu(x)]\psi(x)\}.$$

Here, e and m are charge and mass parameters associated with the Klein–Gordon field and the bar stands for the complex conjugation. In this combined system of quasilinear, second-order partial differential equations, there exists *one* differential identity, viz.

$$\partial_\mu M^\mu + ie[\bar{\psi}(x)k(x) - \bar{k}(x)\psi(x)] \equiv 0. \tag{2.2}$$

This identity holds even if $M^\mu \neq 0, k(x) \neq 0$. In the system of equations (2.1) we have six (real) unknown functions and five (algebraically and differentially) independent equations. Therefore, to make the system determinate, we are entitled to add one subsidiary condition which is not inconsistent with the system (2.1). We choose the subsidiary condition to be the Lorentz-gauge condition:

$$L(x) := \partial_\mu A^\mu(x) = 0. \tag{2.3}$$

The system of equations (2.1), (2.3) is invariant under the *restricted* gauge transformation:

$$\hat{\psi}(x) = \psi(x)e^{ie\lambda(x)},$$
$$\hat{A}_\mu(x) = A_\mu(x) - \partial_\mu\lambda(x), \tag{2.4}$$
$$\Box\lambda := \partial^\mu\partial_\mu\lambda = 0.$$

The symmetrized energy-momentum-stress tensor components for the combined MKG fields are given by

$$\theta_{\mu\nu}(x) := \{[\partial_\mu - ieA_\mu(x)]\bar{\psi}\}\{[\partial_\nu + ieA_\nu(x)]\psi\} + \{[\partial_\nu - ieA_\nu(x)]\bar{\psi}\}\{[\partial_\mu + ieA_\mu(x)]\psi\} - \eta_{\mu\nu}\{[(\partial^\lambda - ieA^\lambda(x))\bar{\psi}][(\partial_\lambda + ieA_\lambda(x))\psi] - m^2|\psi(x)|^2\} - F_{\mu\lambda}(x)F^\lambda{}_\nu(x) + \tfrac{1}{4}\eta_{\mu\nu}F^{\alpha\beta}(x)F_{\alpha\beta}(x). \tag{2.5}$$

There exist four differential identities:

$$\partial_\nu\theta^{\mu\nu} - \bar{k}(x)\{[\partial^\mu + ieA^\mu(x)]\psi\} - k(x) \times\{[\partial^\mu - ieA^\mu(x)]\bar{\psi}\} + F^\mu{}_\nu(x)M^\nu(x) \equiv 0. \tag{2.6}$$

The identities (2.6) are valid even in the case when $k(x) \neq 0$, $M^\nu(x) \neq 0$.

The general solution of a system of partial differential equations contain *arbitrary functions*. In the present system, (2.1), (2.3), the most general solution should contain at least three real-valued and two complex-valued arbitrary functions.

III. PLANE-WAVE SOLUTIONS

Let us assume that the special plane-wave function of the Klein–Gordon field is of the form:

$$\psi(x) = \alpha e^{-ip_\mu x^\mu}, \tag{3.1}$$

where α is an arbitrary complex constant. Assuming that $e \neq 0$, we can define a gauge function $\lambda(x) := p_\mu x^\mu/e$. It satisfies the wave equation $\Box\lambda = 0$ everywhere. Therefore, we are entitled to make a gauge transformation (2.4) on the wave function in (3.1) to obtain:

$$\hat{\psi}(x) = \psi(x)e^{ie\lambda(x)} = (\alpha e^{-ip_\mu x^\mu})e^{ip_\mu x^\mu} = \alpha, \tag{3.2}$$
$$\hat{A}_\mu(x) = A_\mu(x) - p_\mu/e.$$

Dropping the circumflexes in the sequel, the system of equations (2.1), (2.3), with (3.2) reduces to

$$k(x) = [m^2 - e^2A^\mu(x)A_\mu(x) + ie(\partial_\mu A^\mu)]\alpha = 0, \tag{3.3a}$$
$$M^\mu(x) = \Box A^\mu(x) + 2e^2|\alpha|^2A^\mu(x) = 0, \tag{3.3b}$$
$$L(x) = \partial_\mu A^\mu(x) = 0. \tag{3.3c}$$

In case $\alpha = 0$, we have $k(x) \equiv 0$ and (3.3b), (3.3c) reduce to the usual Maxwell equation. However, in case $\alpha \neq 0$, some interesting properties for the combined field emerge from the system (3.3a), (3.3b), (3.3c).

Theorem 3.1: Let $D \subset \mathbb{R}^4$ correspond to a domain of flat space-time M. Let the potential functions $A^\mu \in \mathscr{C}^3(D;\mathbb{R})$ and the field equations (3.3a), (3.3b), (3.3c) hold in D with $m \neq 0, e \neq 0, \alpha \neq 0$. Then (i) there exist two distinct subcases of solutions corresponding to $A^\mu(x)$ being future-pointing timelike or else being a past-pointing timelike vector field; (ii) there exists an algebraic identity $\mathrm{Im}[k(x)/(\alpha e)] - L(x) \equiv 0$; (iii) the charge-current four-vector $j^\mu(x)$ is a constant multiple of $A^\mu(x)$; (iv) the four-potential functions $A^\mu(x)$ satisfy the Proca equations.

Proof: (i) By Eq. (3.3a) we have

$$\mathrm{Re}[k(x)/\alpha] = m^2 - e^2A^\mu(x)A_\mu(x) = 0.$$

Therefore,

$$A^\mu(x)A_\mu(x) = (m/e)^2. \tag{3.4}$$

Since $(m/e)^2 > 0$, $A^\mu(x)$ must be timelike everywhere in D. Solving the quadratic relation (3.4), we obtain two branches (or two distinct functions):

$$A^0(x) = \pm\sqrt{[A^1(x)]^2 + [A^2(x)]^2 + [A^3(x)]^2 + (m/e)^2} \neq 0.$$

Choosing the upper sign, $A^\mu(x)$ is future-pointing everywhere whereas choosing the lower sign $A^\mu(x)$ is past pointing everywhere in D. [$A^0(x)$ is zero nowhere].

(ii) By Eqs. (3.3a) and (3.3c) the algebraic identity

$$\mathrm{Im}[k(x)/(\alpha e)] - L(x) \equiv 0$$

follows immediately.

(iii) The charge-current vector field $j^\mu(x)$ from equations (2.1) and (3.3b) is given by

$$j^\mu(x) = \Box A^\mu(x) = -2e^2|\alpha|^2A^\mu(x).$$

(iv) Equation (3.3b) is equivalent to the Proca equation:

$$\Box A^{\mu}(x) + M^2 A^{\mu}(x) = 0, \tag{3.5}$$

where $M := \sqrt{2}e|\alpha|$. Thus the proof of the theorem is complete. ■

Now we shall attempt to solve the system (3.3a), (3.3b), and (3.3c). To solve the Lorentz gauge condition (3.3c) we introduce the differential one-form:

$$A(x) := A_{\mu}(x)dx^{\mu}.$$

The corresponding dual three-form is

$$A^*(x) = \tfrac{1}{6}\eta_{\mu\nu\lambda\rho}A^{\rho}(x)dx^{\mu}\wedge dx^{\nu}\wedge dx^{\lambda},$$

where $\eta_{\mu\nu\lambda\rho}$ is the totally antisymmetric pseudotensor. The exterior derivative

$$dA^*(x) = -(\partial_{\mu}A^{\mu}(x))dx^0\wedge dx^1\wedge dx^2\wedge dx^3 \equiv 0.$$

Therefore, by the converse of the Poincaré lemma (in a *contractible* domain $D\subset\mathbb{R}^4$), there exists a two-form

$$a(x) = \tfrac{1}{2}a_{\mu\nu}(x)dx^{\mu}\wedge dx^{\nu}$$

such that

$$A^*(x) = da(x) = \tfrac{1}{6}(\partial_{\mu}a_{\nu\lambda}(x) + \partial_{\nu}a_{\lambda\mu}(x) + \partial_{\lambda}a_{\mu\nu}(x))dx^{\mu}\wedge dx^{\nu}\wedge dx^{\lambda}.$$

The above equation implies that

$$A(x) = A^{**}(x) = (da)^*, \tag{3.6}$$

$$A^{\mu}(x) = \partial_{\nu}a^{\mu\nu}.$$

Conversely, $A^{\mu}(x) = \partial_{\nu}a^{\mu\nu}$ [$a^{\nu\mu}(x) = -a^{\mu\nu}(x)$] satisfies (3.3c). Equation (3.6) gives the general solution of (3.3c), where $a^{\mu\nu}$ are arbitrary functions of class $\mathscr{C}^4(D;\mathbb{R})$. Substituting (3.6) into (3.3b) or (3.5) we obtain:

$$\partial_{\nu}[\Box a^{\mu\nu}(x) + M^2 a^{\mu\nu}(x)] = 0. \tag{3.7}$$

Using the techniques of differential forms we can derive

$$\Box a^{\mu\nu}(x) + M^2 a^{\mu\nu}(x) = \eta^{\mu\nu\lambda\rho}(\partial_{\lambda}B_{\rho}(x) - \partial_{\rho}B_{\lambda}(x)), \tag{3.8}$$

where the functions B_{ρ} belong to $\mathscr{C}^5(D;\mathbb{R})$, but otherwise arbitrary. The general solution of (3.8) can be written as

$$a^{\mu\nu}(x) = a^{\mu\nu}_{(h)}(x) + a^{\mu\nu}_{(p)}(x),$$

$$\Box a^{\mu\nu}_{(h)}(x) + M^2 a^{\mu\nu}_{(h)}(x) = 0, \tag{3.9}$$

$$a^{\mu\nu}_{(p)}(x) = \int G(x-x')\eta^{\mu\nu\lambda\rho}(\partial'_{\lambda}B_{\rho} - \partial'_{\rho}B_{\lambda})d^4x'.$$

Here, $a^{\mu\nu}_{(h)}(x)$ is an arbitrary solution of the Proca equation and $G(x-x')$ is an inhomogeneous Green's function for the Proca equation. The remaining equation (3.4) can be expressed as

$$\eta_{\mu\nu}[\partial_{\lambda}(a^{\lambda\nu}_{(h)}(x) + a^{\lambda\nu}_{(p)}(x))][\partial_{\alpha}(a^{\mu\alpha}_{(h)}(x) + a^{\mu\alpha}_{(p)}(x))] = (m/e)^2. \tag{3.10}$$

In this stage, MKG equations have been reduced to *one nonlinear equation* (3.10). Equation (3.10) involves the four functions B_{λ} and six Proca functions $a^{\mu\nu}_{(h)}(x)$. Therefore, Eq. (3.10) is highly underdetermined. Nevertheless, it is not easy to integrate (3.10) in general.

IV. STEADY-STATE FIELD EQUATIONS

We shall define, in general, the steady state of stationary state of the MKG fields by the following criteria:

$$\partial_0 j^{\mu} \equiv 0, \tag{4.1a}$$

$$\partial_0 \theta^{\mu\nu} \equiv 0. \tag{4.1b}$$

In case of the field equations (3.3a), (3.3b), the condition (4.1a) implies that

$$\partial_0 A^{\mu} \equiv 0. \tag{4.2}$$

Let us denote a spatial point by $\mathbf{x} = (x^1,x^2,x^3)$ in a domain $\mathbf{D}\subset\mathbb{R}^3$. The field equations (3.3a), (3.3b), (3.3c) by (4.2) reduce to

$$[A^0(\mathbf{x})]^2 = \delta_{jk}A^j(\mathbf{x})A^k(\mathbf{x}) + (m/e)^2, \tag{4.3a}$$

$$\nabla^2 A^0(\mathbf{x}) - M^2 A^0(\mathbf{x}) = 0, \tag{4.3b}$$

$$\nabla^2 A^k(\mathbf{x}) - M^2 A^k(\mathbf{x}) = 0, \tag{4.3c}$$

$$\partial_k A^k(\mathbf{x}) = 0. \tag{4.3d}$$

Now we shall prove a theorem on the above system of equations in case of the absence of any magnetic field.

Theorem 4.1: Let $\mathbf{D}\subset\mathbb{R}^3$ correspond to a contractible domain of the spatial submanifold of M. Let $A^{\mu}\in\mathscr{C}^3(\mathbf{D};\mathbb{R})$ and the steady-state field equations (4.3a)–(4.3d) hold in $\mathbf{D}$ with $M\neq 0$, $e\neq 0$, $m\neq 0$. If, furthermore, $F_{ij}(\mathbf{x})\equiv 0$ in $\mathbf{D}$, then there exists no solution of the system (4.3a)–(4.3d) in $\mathbf{D}$.

Proof: Since $A^i\in\mathscr{C}^3(\mathbf{D};\mathbb{R})$ and $\mathbf{D}$ is contractible, the equations $F_{ij}(\mathbf{x}) = \partial_j A_i - \partial_i A_j \equiv 0$ imply the existence of a function $\lambda\in\mathscr{C}^4(\mathbf{D};\mathbb{R})$ such that $A_i(\mathbf{x}) = \partial_i\lambda$. Equation (4.3d) reduces to $\nabla^2\lambda = 0$. Therefore, the equations (4.3c) yield $M^2\partial^k\lambda = 0$. Since $M\neq 0$, we must have $A^k(\mathbf{x}) = \partial^k\lambda \equiv 0$. Thus the equations (4.3a) gives $A^0(\mathbf{x}) = \pm(m/e)$. Substituting this result into (4.3b) yields $\mp M^2(m/e) = 0$. This condition contradicts the hypotheses that $M\neq 0$ and $m\neq 0$. ■

It is clear from Theorem 3.1 that the charge-current vector field $j^{\mu}(\mathbf{x})$ can be either future-pointing timelike or else past-pointing timelike. Let us choose $j^{\mu}(\mathbf{x})$ to be future pointing in $\mathbf{D}$. We shall prove another theorem about the consequences of this implication.

Theorem 4.2: Let $\mathbf{D}\subset\mathbb{R}^3$ correspond to a bounded domain of the spatial submanifold of M. Let $A^{\mu}\in\mathscr{C}^3(\mathbf{D};\mathbb{R})$ and the steady-state field equations (4.3a)–(4.3d) hold in $\mathbf{D}$ with $M\neq 0$, $e\neq 0$. Let $j^{\mu}(\mathbf{x})$ be future pointing in $\mathbf{D}$. If furthermore, $A^0(\mathbf{x})$ attains a minimum at some point $\mathbf{x}_m\in\mathbf{D}$ so that for all $\mathbf{x}\in\mathbf{D}$, $A^0(\mathbf{x}_m)\leqslant A^0(\mathbf{x})$, then the system (4.3a)–(4.3d) has no solution in $\mathbf{D}$.

Proof: When $j^{\mu}(\mathbf{x})$ is future pointing, then $j^0(\mathbf{x}) > 0$. Therefore, $A^0(\mathbf{x}) = -M^{-2}j^0(\mathbf{x}) < 0$. By Eq. (4.3b) it follows that $\nabla^2 A^0(\mathbf{x}) = M^2 A^0(\mathbf{x}) < 0$. As a consequence of $A^{\mu}\in\mathscr{C}^3(\mathbf{D};\mathbb{R})$ and Hopf's theorem[5] on elliptic differential inequalities, it follows that $A^0(\mathbf{x})$ is constant valued and $A^0(\mathbf{x}) = A^0(\mathbf{x}_m)$. By Eq. (4.3b) and $M\neq 0$, we must have $A^0(\mathbf{x})\equiv 0$ for all $\mathbf{x}\in\mathbf{D}$. Therefore, $j^0(\mathbf{x}) = -M^2A^0(\mathbf{x})\equiv 0$, which contradicts $j^0(\mathbf{x}) > 0$. ■

V. A SPECIAL CLASS OF STEADY-STATE SOLUTIONS

We shall investigate the steady-state field equations (4.3a)–(4.3d) under the special assumption that A^μ depend on one independent variable. In case

$$\partial_1 A^\mu = \partial_2 A^\mu \equiv 0, \tag{5.1}$$

Eq. (4.3d) implies that A^3 is constant valued. Therefore, the field equation $\nabla^2 A^3 - M^2 A^3(\mathbf{x}) = 0$ with $M \neq 0$ can be satisfied only by the choice

$$A^3(\mathbf{x}) \equiv 0. \tag{5.2}$$

Equations (4.3b) and (4.3c) can be integrated to obtain

$$\begin{aligned} A^0(\mathbf{x}) &= a_1 e^{Mx^3} + a_2 e^{-Mx^3}, \\ A^1(\mathbf{x}) &= b_1 e^{Mx^3} + b_2 e^{-Mx^3}, \\ A^2(\mathbf{x}) &= c_1 e^{Mx^3} + c_2 e^{-Mx^3}, \end{aligned} \tag{5.3}$$

where $a_1, a_2, b_1, b_2, c_1, c_2$ are *arbitrary* constants of integration. Using (5.2), (5.3) Eq. (4.3a) yields

$$\begin{aligned} &2[a_1 a_2 - b_1 b_2 - c_1 c_2 - \tfrac{1}{2}(m/e)^2] \\ &\quad + (a_1^2 - b_1^2 - c_1^2) e^{2Mx^3} \\ &\quad + (a_2^2 - b_2^2 - c_2^2) e^{-2Mx^3} \equiv 0. \end{aligned} \tag{5.4}$$

Since the set of functions $\{1, e^{2Mx^3}, e^{-2Mx^3}\}$ is linearly independent on an interval, the identity (5.4) implies that

$$\begin{aligned} &a_1 a_2 - b_1 b_2 - c_1 c_2 = (m/e)^2/2, \\ &a_1^2 - b_1^2 - c_1^2 = 0, \\ &a_2^2 - b_2^2 - c_2^2 = 0. \end{aligned} \tag{5.5}$$

The general solution of the nonlinear relations (5.5) is as follows:

$$\begin{aligned} a_2 &= (m/e)^2 \{2a_1 [1 - \cos(\phi_1 - \phi_2)]\}^{-1}, \\ b_1 &= a_1 \cos\phi_1, \\ c_1 &= a_1 \sin\phi_1, \\ b_2 &= a_2 \cos\phi_2, \\ c_2 &= a_2 \sin\phi_2, \end{aligned} \tag{5.6}$$

where $a_1 \neq 0$, and $\phi_1 - \phi_2 \neq 2n\pi$, but otherwise a_1, ϕ_1, and ϕ_2 are arbitrary real constants. The electromagnetic four-potential components are

$$\begin{aligned} A_0(\mathbf{x}) &= a_1 e^{Mx^3} + a_2 e^{-Mx^3}, \\ A_1(\mathbf{x}) &= -(a_1 \cos\phi_1 e^{Mx^3} + a_2 \cos\phi_2 e^{-Mx^3}), \\ A_2(\mathbf{x}) &= -(a_1 \sin\phi_1 e^{Mx^3} + a_2 \sin\phi_2 e^{-Mx^3}), \\ A_3(\mathbf{x}) &\equiv 0. \end{aligned} \tag{5.7}$$

The electromagnetic field tensor components and charge-current vector components are

$$\begin{aligned} F_{12}(\mathbf{x}) &= F_{01}(\mathbf{x}) = F_{02}(\mathbf{x}) \equiv 0, \\ F_{31}(\mathbf{x}) &= M(a_1 \cos\phi_1 e^{Mx^3} - a_2 \cos\phi_2 e^{-Mx^3}), \\ F_{23}(\mathbf{x}) &= -M(a_1 \sin\phi_1 e^{Mx^3} - a_2 \sin\phi_2 e^{-Mx^3}), \\ F_{03}(\mathbf{x}) &= M(a_1 e^{Mx^3} - a_2 e^{-Mx^3}), \\ F_{\mu\nu}(\mathbf{x}) F^{\mu\nu}(\mathbf{x}) &= 4M^2|\alpha|^2, \\ j^0(\mathbf{x}) &= -M^2(a_1 e^{Mx^3} - a_2 e^{-Mx^3}), \end{aligned} \tag{5.8}$$

$$\begin{aligned} j^1(\mathbf{x}) &= -M^2(a_1 \cos\phi_1 e^{Mx^3} + a_2 \cos\phi_2 e^{-Mx^3}), \\ j^2(\mathbf{x}) &= -M^2(a_1 \sin\phi_1 e^{Mx^3} + a_2 \sin\phi_2 e^{-Mx^3}), \\ j^3(\mathbf{x}) &\equiv 0, \\ \operatorname{sgn}(\rho(\mathbf{x})) &= -\operatorname{sgn}(a_1). \end{aligned}$$

Using Eqs. (2.5), (5.7), some of the components of energy-momentum-stress tensor of MKG fields are

$$\begin{aligned} \theta_{00}(\mathbf{x}) &= M^2|\alpha|^2 + 2M^2(a_1^2 e^{2Mx^3} + a_2^2 e^{-2Mx^3}) > 0, \\ \theta_{10}(\mathbf{x}) &= -2M^2(a_1^2 \cos\phi_1 e^{2Mx^3} + a_2^2 \cos\phi_2 e^{-2Mx^3}), \\ \theta_{20}(\mathbf{x}) &= -2M^2(a_1^2 \sin\phi_1 e^{2Mx^3} + a_2^2 \sin\phi_2 e^{-2Mx^3}), \\ \theta_{30}(\mathbf{x}) &\equiv 0, \\ &[\theta_{00}(\mathbf{x}) - M^2|\alpha|^2]^2 - [\theta_{10}(\mathbf{x})]^2 \\ &\quad - [\theta_{20}(\mathbf{x})]^2 - [\theta_{30}(\mathbf{x})]^2 \\ &= 8(m|\alpha|)^4 [1 - \cos(\phi_1 - \phi_2)]^{-1}. \end{aligned} \tag{5.9}$$

Let us choose a bounded *cylindrical* domain $\mathbf{D} = B \times (h_1, h_2)$, where the base $B \subset \mathbb{R}^2$. In such a case, the *total* energy and momentum components of the coupled MKG fields inside the domain $\mathbf{D}$ are given by

$$\begin{aligned} P_0 &:= \int_{\mathbf{D}} \theta_{00}(\mathbf{x}) d^3\mathbf{x} \\ &= \{M^2|\alpha|^2 + M(h_2 - h_1)^{-1}[a_1^2(e^{2Mh_2} - e^{2Mh_1}) \\ &\quad - a_2^2(e^{-2Mh_2} - e^{-2Mh_1})]\} V(\mathbf{D}), \\ P_1 &:= \int_{\mathbf{D}} \theta_{10}(\mathbf{x}) d^3\mathbf{x} \\ &= -M(h_2 - h_1)^{-1}[a_1^2 \cos\phi_1 (e^{2Mh_2} - e^{2Mh_1}) \\ &\quad - a_2^2 \cos\phi_2 (e^{-2Mh_2} - e^{-2Mh_1})] V(\mathbf{D}), \\ P_2 &:= \int_{\mathbf{D}} \theta_{20}(\mathbf{x}) d^3\mathbf{x} \\ &= -M(h_2 - h_1)^{-1}[a_1^2 \sin\phi_1 (e^{2Mh_2} - e^{2Mh_1}) \\ &\quad - a_2^2 \sin\phi_2 (e^{-2Mh_2} - e^{-2Mh_1})] V(\mathbf{D}), \\ P_3 &:= \int_{\mathbf{D}} \theta_{30}(\mathbf{x}) d^3\mathbf{x} \equiv 0, \end{aligned}$$

where $V(\mathbf{D})$ stands for the volume of the cylindrical domain $\mathbf{D}$. The above integrals usually diverge for an unbounded domain. The total mass in the cylinder may be represented by

$$\sqrt{\eta_{\mu\nu} P^\mu P^\nu} = F(h_1, h_2, \phi_1, \phi_2) V(\mathbf{D}),$$

where F is a known complicated function.

[1] Y. Choquet-Bruhat, C. R. Acad. Sci. (Paris) Ser. I **292**, 153 (1981); M. Flato, J. Simon, and E. Taflin, Commun. Math. Phys. **112**, 21 (1987).
[2] A. Das and D. Kay, J. Math. Phys. **30**, 2280 (1989).
[3] A. Das, J. Math. Phys. **4**, 45 (1963), A. Das and C. V. Coffmann, J. Math. Phys. **8**, 1720 (1967), J. Gegenberg and A. Das, J. Math. Phys. **22**, 1736 (1981).
[4] D. Brill and S. Cohen, J. Math. Phys. **7**, 236 (1966).
[5] K. Yano and S. Bochner, *Curvature and Betti Numbers* (Princeton U. P., Princeton, NJ, 1953), p. 26.

The functional Ito formula in quantum stochastic calculus

G. O. S. Ekhaguere
Department of Mathematics, University of Ibadan, Ibadan, Nigeria

(Received 3 July 1986; accepted for publication 6 June 1990)

Using an Op-*-algebraic approach, noncommutative analogs of the Ito formula of classical stochastic calculus within the framework of the Hudson–Parthasarathy formulation of Boson quantum stochastic calculus are proven.

I. INTRODUCTION

In classical probability theory, the stochastic integral, first defined by Ito,[1] and extended by Kunita and Watanabe[2] and by Meyer,[3,4] plays an important role. As a result of that notion of integration, it has been possible, for example, to discuss stochastic differential equations and diffusion processes[5–7] that are significant in the understanding of various natural phenomena. It is, therefore, interesting that, in recent times, several publications dealing with noncommutative extensions[8–10] of classical stochastic integration have appeared.

In this paper, we present results that are the noncommutative analogs of the well-known Ito formula[11] for the Boson quantum stochastic integral introduced by Hudson and Parthasarathy.[10] Our algebraic approach enables us to include unbounded linear operators as integrands, in distinction to the consideration in Ref. 10.

This contribution is organized as follows. In Sec. II, we present the fundamental concepts and notation that feature in the subsequent discussion. We introduce Op-*-algebras[12] of unbounded linear operators on certain Hilbert spaces and consider some locally convex completions of the algebras in Sec. III. The complete locally convex spaces are employed in Sec. IV where we present noncommutative versions of the Ito formula.[11] We note that extensions of the Ito formula have appeared in various contexts such as those concerning: Hilbert-space-valued processes;[13] Banach-space-valued processes;[14] processes in abstract Wiener spaces;[15] processes in infinite dimensional manifolds;[16] operator-valued processes;[17,18] and quantum stochastic processes.[10,19,20] We remark that the noncommutative extension of the Ito formula presented in Ref. 10. Theorem (4.5) is a special case of Theorem (4.10) of this paper.

II. FUNDAMENTAL CONCEPTS AND NOTATION

If $\mathscr{L}$ is a Hilbert space, then $\|\cdot\|_{\mathscr{L}}$ and $\langle .,. \rangle_{\mathscr{L}}$ denote its norm and inner product, respectively. Furthermore, the following notation is used throughout the paper:

ds = Lebesgue measure;

I = any Borel subset of $\mathbb{R}_+ \equiv [0,\infty)$;

$\mathscr{K}$ = some fixed Hilbert space;

$L^2_{\mathscr{K}}(I,ds)$ = set of all (equivalence classes) of $\mathscr{K}$-valued functions f on I such that $\int_I ds\|f(s)\|^2_{\mathscr{K}} < \infty$;

$\mathscr{H}(I) = L^2_{\mathscr{K}}(I,ds)$;

$\mathscr{H} = \mathscr{H}(\mathbb{R}_+)$;

$B(\mathscr{H}(I))$ = Banach space of all endomorphisms of $\mathscr{H}(I)$;

$L^{\infty,\mathrm{loc}}_{\mathscr{H}(I)}(I)$ = linear space of all $\mathscr{H}(I)$-valued, locally bounded, Lebesgue measurable functions on I;

$L^{\infty,\mathrm{loc}}_{B(\mathscr{H}(I))}(I)$ = linear space of all $B(\mathscr{H}(I))$-valued, locally bounded, Lebesgue measurable functions on I; the members of $L^{\infty,\mathrm{loc}}_{B(\mathscr{H}(I))}(I)$ act pointwise on $\mathscr{H}(I)$;

$\Gamma(\mathscr{H}(I))$ = Boson Fock space[21,22] over $\mathscr{H}(I)$.

It is well known[22] that the exponential vectors in $\Gamma(\mathscr{H}(I))$ generate a dense linear subspace. Recall that an exponential vector is of the form:

$$\Omega(f) = \bigoplus_{n=0}^{\infty} (n!)^{-1/2}(f\otimes f\otimes\cdots\otimes f)_n, \quad f\in\mathscr{H}(I),$$

where $(f\otimes f\otimes\cdots\otimes f)_0 \equiv 1$ and $(f\otimes f\otimes\cdots\otimes f)_n$ is the n-fold tensor product of f with itself.

For $f\in\mathscr{H}(I)$ and $\pi\in L^{\infty,\mathrm{loc}}_{B(\mathscr{H}(I))}(I)$, introduce the linear operators $a(f)$, $a^*(f)$, and $\lambda(\pi)$, defined on the linear span $V\{\Omega(f): f\in\mathscr{H}(I)\}$, as follows:

$$a(f)\Omega(g) = \langle f,g\rangle_{\mathscr{H}(I)}\Omega(g),$$

$$a^*(f)\Omega(g) = \frac{d}{dx}\Omega(g+xf)|_{x=0},$$

$$\lambda(\pi)\Omega(g) = \frac{d}{dx}\Omega(e^{x\pi}g)_{x=0}, \quad g\in\mathscr{H}(I).$$

Observe that $a(f)$ and $a^*(f)$, $f\in\mathscr{H}(I)$, are the annihilation and creation operators of quantum field theory;[21] $\lambda(\pi)$ is the so-called gauge operator.

It is noteworthy that the linear span of the exponential vectors is not left invariant under the action of the operators $a^*(f)$ and $\lambda(\pi)$. But in the subsequent discussion, we need a dense domain that is left invariant under the action of arbitrary polynomials in the operators $a(f)$, $a^*(f)$, and $\lambda(\pi)$ for any $f\in\mathscr{H}(I)$ and $\pi\in L^{\infty,\mathrm{loc}}_{B(\mathscr{H}(I))}(I)$. Such an invariant subspace of $\Gamma(\mathscr{H}(I))$ is obtained as follows.

Let $\mathfrak{S}(I)$ be the linear subspace of $\Gamma(\mathscr{H}(I))$ generated by

$$\{\Omega(f): f\in\mathscr{H}(I)\}$$

and

$$\left\{\frac{\partial^{n+m}}{\partial x^n\,\partial y^n}\Omega(f+xg+e^{y\pi}h)\Big|_{\substack{x=0\\y=0}}: f,g,h\in\mathscr{H}(I),\ \pi\in L^{\infty,\mathrm{loc}}_{B(\mathscr{H}(I))}(I),\ n,m=1,2,\ldots\right\}.$$

Then, $\mathfrak{S}(I)$ is a dense subspace of $\Gamma(\mathscr{H}(I))$ which is invar-

 0022-2488/90/122921-09$03.00

iant under the action of any member of the polynomial algebra generated by

$$\{a(f),a^*(g),\lambda(\pi):f,g,\in\mathscr{H}(I),\pi\in L^{\infty,\mathrm{loc}}_{B(\mathscr{H}(I))}(I)\}.$$

For $f\in\mathscr{H}(I)$ and $\pi\in L^{\infty,\mathrm{loc}}_{B(\mathscr{H}(I))}(I)\}$ the closures of $a(f)$, $a^*(f)$, and $\lambda(\pi)$ (denoted again by the same symbols) act on $\mathfrak{S}(I)$. We set

$$\mathfrak{S}(\mathbb{R}_+)\equiv\mathfrak{S};\quad \mathfrak{S}([0,t))\equiv\mathfrak{S}_t;\quad \mathfrak{S}([t,\infty))\equiv\mathfrak{S}^t,\quad t\in\mathbb{R}_+.$$

III. QUANTUM Op-*-ALGEBRAS AND ASSOCIATED LOCALLY CONVEX SPACES

We employ Op-*-algebras in the sequel. Those algebras and some associated spaces are introduced in this section.

A. Op-*-algebras

If $\mathscr{F}$ is some dense subspace of a Hilbert space $\mathscr{G}$, with the inner product $\langle\cdot,\cdot\rangle_{\mathscr{G}}$, we write $\mathfrak{A}(\mathscr{F})$ for the set of all linear operators that leave $\mathscr{F}$ invariant and possess adjoints whose domains contain $\mathscr{F}$. Then, $\mathfrak{A}(\mathscr{F})$ is a unital *-algebra, called an Op-*-algebra,[12] when equipped with the involution $x\to x^*$ defined by $\langle\xi_1,x\xi_2\rangle_{\mathscr{G}}=\langle x^*\xi_1,\xi_2\rangle_{\mathscr{G}}$, $\xi_1,\xi_2\in\mathscr{F}$. We write $\mathbf{1}$ for the unit of $\mathfrak{A}(\mathscr{F})$.

In the sequel, $\mathscr{R}$ always denotes some fixed Hilbert space and $\mathscr{K}$ is the Hilbert space:

$$\mathscr{K}\equiv\mathscr{R}\otimes\Gamma(\mathscr{H}).$$

Since $\mathscr{H}=\mathscr{H}([0,t))\oplus\mathscr{H}([t,\infty))$, for each $t\in\mathbb{R}_+$, it follows that $\mathscr{K}$ admits the factorization:

$$\mathscr{K}=\mathscr{K}_t\otimes\Gamma(\mathscr{H}([t,\infty)))\quad\text{for each } t\in\mathbb{R}_+,$$

where

$$\mathscr{K}_t=\mathscr{R}\otimes\Gamma(\mathscr{H}([0,t))),\quad t\in\mathbb{R}_+.$$

The space $\mathscr{K}_0$ is identified with $\mathscr{R}$ and called the *initial space.*

Throughout this paper, $\mathscr{D}$ denotes a dense subspace of $\mathscr{R}$. Then $\mathscr{E}\equiv\mathscr{D}\otimes\mathfrak{S}$ is dense in $\mathscr{K}$. Moreover, $\mathscr{E}_t\equiv\mathscr{D}\otimes\mathfrak{S}_t$ and $\mathfrak{S}^t$ are dense in $\mathscr{K}_t$ and $\Gamma(\mathscr{H})([t,\infty)))$, respectively, for each $t\in\mathbb{R}_+$. Hence, the Op-*-algebras $\mathfrak{A}(\mathscr{E}_t)\otimes\mathbb{1}^t$ and $\mathbf{1}_t\otimes\mathfrak{A}(\mathfrak{S}^t)$ may be identified with Op-*-subalgebras of $\mathfrak{A}(\mathscr{E})$, where $\mathbf{1}_t$ and $\mathbf{1}^t$ are the units in $\mathfrak{A}(\mathscr{E}_t)$ and $\mathfrak{A}(\mathfrak{S}^t)$, respectively. In the sequel, we set $\mathfrak{A}(\mathscr{E})\equiv\mathfrak{A}$; $\mathfrak{A}(\mathscr{E}_t)\otimes\mathbb{1}^t\equiv\mathfrak{A}_t$; and $\mathbf{1}_t\otimes\mathfrak{A}(\mathfrak{S}^t)\equiv\mathfrak{A}^t$.

B. The space $\mathfrak{A}(\tau_2)$

For $\alpha\in\mathscr{D}\times\mathscr{H}$, with $\alpha=(u,f)$, let $\|\cdot\|_\alpha$ be the semi-norm on $\mathfrak{A}$ defined as follows:

$$\|b\|_\alpha=\|bu\otimes\Omega(f)\|_{\mathscr{K}},\quad b\in\mathfrak{A}.$$

Then, we write τ_2 for the locally convex topology generated by $\{\|\cdot\|_\alpha:\alpha\in\mathscr{D}\times\mathscr{H}\}$ and denote the τ_2-completion of $\mathfrak{A}$, $\mathfrak{A}_t$, and $\mathfrak{A}^t$, $t\in\mathbb{R}_+$, by $\mathfrak{A}(\tau_2)$, $\mathfrak{A}_t(\tau_2)$, and $\mathfrak{A}^t(\tau_2)$, respectively.

In the sequel, $\mathrm{Fin}(\mathscr{D}\times\mathscr{H})$ denotes the collection of all nonempty *finite* subsets of $\mathscr{D}\times\mathscr{H}$. For each $\mathbb{H}\in\mathrm{Fin}(\mathscr{D}\times\mathscr{H})$, we define $\|\cdot\|_{\mathbb{H}}$ on $\mathfrak{A}(\tau_2)$ as follows:

$$\|x\|_{\mathbb{H}}=\sup_{\alpha\in\mathbb{H}}\|x\|_\alpha,\quad x\in\mathfrak{A}(\tau_2).$$

C. The space $\mathscr{L}^{(0)}_b((\mathfrak{A}(\tau_2))^n,\mathscr{B})$

Let $\mathscr{B}$ be a complete locally convex space whose topology is generated by the family $\{|||\cdot|||_\delta,\delta\in\Delta\}$ of semi-norms. Then, $\mathscr{L}(\mathfrak{A}(\tau_2),\mathscr{B})$ denotes the collection of all *continuous* linear mappings from $\mathfrak{A}(\tau_2)$ into $\mathscr{B}$. For $\eta\in\mathscr{L}(\mathfrak{A}(\tau_2),\mathscr{B})$, $\eta[x]$ denotes the value of η at the point $x\in\mathfrak{A}(\tau_2)$.

Let $\eta\in\mathscr{L}(\mathfrak{A}(\tau_2),\mathscr{B})$. Then, by [Ref. 23, Theorem III.1.1], for each $\delta\in\Delta$, there exist a positive number $c_{\delta,\eta}$ and a member $\mathbb{H}(\delta,\eta)$ of $\mathrm{Fin}(\mathscr{D}\times\mathscr{H})$ such that

$$(*)\quad |||\eta[x]|||_\delta\leqslant c_{\delta,\eta}\|x\|_{\mathbb{H}(\delta,\eta)},\quad\text{for all } x\in\mathfrak{A}(\tau_2).$$

In the sequel, $\mathbb{H}(\cdot)$ is a fixed, but otherwise arbitrary, map from Δ into $\mathrm{Fin}(\mathscr{D}\times\mathscr{H})$. Using (*), we introduce the subset $\mathscr{L}^{(0)}(\mathfrak{A}(\tau_2),\mathscr{B})$ of $\mathscr{L}(\mathfrak{A}(\tau_2),\mathscr{B})$ as follows.

Definition: Let $\eta\in\mathscr{L}(\mathfrak{A}(\tau_2),\mathscr{B})$. Then, we say that η lies in $\mathscr{L}^0(\mathfrak{A}(\tau_2),\mathscr{B})$ if for each $\delta\in\Delta$, we have $|||\eta[x]|||_\delta\leqslant c\|x\|_{\mathbb{H}(\delta)}$ for all $x\in\mathfrak{A}(\tau_2)$, where c is some positive constant depending on δ, η, $\mathbb{H}(\cdot)$.

Remarks: (1) The set $\mathscr{L}^{(0)}(\mathfrak{A}(\tau_2),\mathscr{B})$ is a vector space.

(2) We shall next equip $\mathscr{L}^{(0)}(\mathfrak{A}(\tau_2),\mathscr{B})$ with a topology.

Let $\mathscr{A}$ be a family of bounded subsets of $\mathfrak{A}(\tau_2)$ such that the linear hull of $\bigcup_{\Theta\in\mathscr{A}}\{x\in\Theta\}$ is dense in $\mathfrak{A}(\tau_2)$. Then, for each $(\delta,\Theta)\in\Delta\times\mathscr{A}$, define $|||\cdot|||_{\delta,\Theta}$ on $\mathscr{L}^{(0)}(\mathfrak{A}(\tau_2),\mathscr{B})$ as follows:

$$|||\eta|||_{\delta,\Theta}=\inf\{c:|||\eta[x]|||_\delta\leqslant c\|x\|_{\mathbb{H}(\delta)},\forall\mathscr{P}\in\Theta\},$$

$$\eta\in\mathscr{L}^{(0)}(\mathfrak{A}(\tau_2),\mathscr{B}).$$

The completion of $\mathscr{L}^{(0)}(\mathfrak{A}(\tau_2),\mathscr{B})$ in the locally convex topology [which we call the $(\mathscr{A})$-topology in the sequel] determined by the family

$$\{|||\cdot|||_{\delta,\Theta}:(\delta,\Theta)\in\Delta\times\mathscr{A}\}$$

of semi-norms will be denoted by $\mathscr{L}^{(0)}_{(\mathscr{A})}(\mathfrak{A}(\tau_2),\mathscr{B})$.

If, for $j=1,2,\ldots,n$, $\mathscr{A}_j$ is a family, as above of bounded subsets of $\mathfrak{A}(\tau_2)$ such that $\bigcup_{\Theta\in\mathscr{A}_j}\{x\in\Theta\}$ is dense in $\mathfrak{A}(\tau_2)$, then we define

$$\mathscr{L}^{(0)}_{(\mathscr{A}_1\times\cdots\times\mathscr{A}_j)}((\mathfrak{A}(\tau_2))^j,\mathscr{B})=\mathscr{L}^{(0)}_{(\mathscr{A}_j)}(\mathfrak{A}(\tau_2),\mathscr{L}^{(0)}_{(\mathscr{A}_1\times\cdots\times\mathscr{A}_{j-1})}((\mathfrak{A}(\tau_2))^{j-1},\mathscr{B})),\quad j=1,2,\ldots,n,$$

with $\mathscr{L}^{(0)}_{(\mathscr{A}_0)}((\mathfrak{A}(\tau_2))^0,\mathscr{B})\equiv\mathscr{B}$.

It is then evident that $\mathscr{L}^{(0)}_{(\mathscr{A}_1\times\cdots\times\mathscr{A}_n)}((\mathfrak{A}(\tau_2))^n,\mathscr{B})$ is isomorphic to the linear space, denoted again by $\mathscr{L}^{(0)}_{(\mathscr{A}_1\times\cdots\times\mathscr{A}_n)}((\mathfrak{A}(\tau_2))^n,\mathscr{B})$, of all the multilinear mappings of $(\mathfrak{A}(\tau_2))^n$ into $\mathscr{B}$ that are separately continuous in the $(\mathscr{A}_1\times\cdots\times\mathscr{A}_n)$-topology.

If $\mathscr{A}_1=\cdots=\mathscr{A}_n\equiv\mathscr{A}$ and $\mathscr{A}$ consists of *all the bounded subsets* of $\mathfrak{A}(\tau_2)$, then we denote $\mathscr{L}^{(0)}_{(\mathscr{A}\times\cdots\times\mathscr{A})}((\mathfrak{A}(\tau_2))^n,\mathscr{B})$ simply by $\mathscr{L}^{(0)}_b((\mathfrak{A}(\tau_2))^n,\mathscr{B})$. Furthermore, we write $\mathscr{L}^{(0)}_b(\mathfrak{A}(\tau_2),\mathfrak{A}(\tau_2))$ as $\mathscr{L}^{(0)}_b(\mathfrak{A}(\tau_2))$.

Let us now conclude this section by introducing the notion of differentiation that we employ in this paper.

Definition: Let $C(\mathfrak{A}(\tau_2))$ be the linear space of all continuous mappings of $\mathfrak{A}(\tau_2)$ into itself. Then by the derivative (if it exists) of $\psi\in C(\mathfrak{A}(\tau_2))$ we mean a map ψ' from $\mathfrak{A}(\tau_2)$ into $\mathscr{L}^{(0)}_b(\mathfrak{A}(\tau_2))$ such that the expression

$(1/s)(\psi(y+sz)-\psi(y))-\psi'(y)[z] \quad (0\neq s\in(-\infty,\infty))$

converges, for each $y\in\mathfrak{A}(\tau_2)$, to the zero member of $\mathfrak{A}(\tau_2)$ as s tends to zero, uniformly with respect to z in any bounded subset of $\mathfrak{A}(\tau_2)$.

Remark: (i) It is evident that any derivative is unique.

(ii) Higher-order derivatives of members of $C(\mathfrak{A}(\tau_2))$ are defined analogously by iteration.

(iii) In the sequel, $C^m(\mathfrak{A}(\tau_2))$ is the linear space of all members of $C(\mathfrak{A}(\tau_2))$ that are m-times differentiable.

(iv) If $\psi\in C^m(\mathfrak{A}(\tau_2))$, we write $\psi^{(m)}$ for its mth derivative. Then, $\psi^{(m)}$ is a map from $\mathfrak{A}(\tau_2)$ into $\mathcal{L}_b^{(0)}((\mathfrak{A}(\tau_2))^m, \mathfrak{A}(\tau_2))$.

Notation: The symbol $C^{1,2}(\mathbf{R}_+\times\mathfrak{A}(\tau_2),\mathfrak{A}(\tau_2))$ denotes the linear space of all continuous mappings ψ from $\mathbf{R}_+\times\mathfrak{A}(\tau_2)$ into $\mathfrak{A}(\tau_2)$ with the properties that (1) for each $y\in\mathfrak{A}(\tau_2)$, the map $t\to\psi(t,y)$ is differentiable in $\mathfrak{A}(\tau_2)$; and (2) for each $t\in\mathbf{R}_+$, the map $y\to\psi(t,y)$ lies in $C^2(\mathfrak{A}(\tau_2))$.

Remark: In the sequel, differentiation with respect to $t\in\mathbf{R}_+$ will be denoted by a dot.

IV. THE ITO FORMULA

In this section, we present results that are the noncommutative generalizations of the *Ito formula*[11] of classical probability theory. To this end, we utilize the theory of stochastic integration expounded by Hudson and Parthasarathy in Ref. 10.

We employ the following notions concerning $\mathfrak{A}$-valued stochastic processes over $\mathbf{R}_+$.

Definition: A map $b:\mathbf{R}_+\to\mathfrak{A}$ is called an *adapted process* if $b(t)\in\mathfrak{A}_t$, for $t\in\mathbf{R}_+$.

Remark: We denote the set of all adapted processes by $\mathrm{Ad}(\mathfrak{A})$. It is clear that $\mathrm{Ad}(\mathfrak{A})$ is a *-algebra when endowed with the operations of pointwise addition, multiplication, scalar multiplication, and involution.

Definition: A member b of $\mathrm{Ad}(\mathfrak{A})$ is called the following.

(1) *simple*, in case b has a representation of the form

$$b(\cdot)=\sum_{n=0}^{\infty}\chi_{(t_n,t_{n+1}]}(\cdot)b_n,$$

where $0=t_0<t_1<\cdots<t_n$, with $t_n\to\infty$ as $n\to\infty$, $b_n\in\mathfrak{A}_{t_n}$, and χ_I is the indicator function of $I\subset\mathbf{R}_+$;

(2) *τ_2-continuous*, in case the map $t\to b(t)$ is continuous from $\mathbf{R}_+$ into $\mathfrak{A}(\tau_2)$; and

(3) *τ_2-locally square integrable*, in case the map $t\to\|b(t)\|_\alpha$, from $\mathbf{R}_+$ into itself, is Lebesgue measurable for each $\alpha\in\mathscr{D}\times\mathscr{H}$ and

$$\int_0^t ds\|b(s)\|_\alpha^2<\infty, \quad \text{for each } t>0.$$

Notation: (1) The collection of all the members of $\mathrm{Ad}(\mathfrak{A})$ that are *simple* (resp. *τ_2-continuous*) will be denoted by $\mathrm{Ad}(\mathfrak{A})_{\mathrm{sim}}$ (resp. $\mathrm{Ad}(\mathfrak{A})_{\mathrm{con}}$).

(2) We write $\mathrm{Ad}(\mathfrak{A},\tau_\infty)$ for the completion of $\mathrm{Ad}(\mathfrak{A})_{\mathrm{con}}$ in the locally convex topology τ_∞ generated by the family

$$\{\|\cdot\|_{\alpha,t,\infty}:\alpha\in\mathscr{D}\times\mathscr{H},\ 0<t\in\mathbf{R}_t\}$$

of semi-norms, where

$$\|b\|_{\alpha,t,\infty}=\sup_{0\le s\le t}\|b(s)\|_\alpha, \quad 0<t\in\mathbf{R}_+, \quad b\in\mathrm{Ad}(\mathfrak{A})_{\mathrm{con}}.$$

(3) The symbol $L^2(\mathfrak{A},\tau_{\mathrm{loc}})$ denotes the completion of the vector space of all *τ_2-locally square integrable* members of $\mathrm{Ad}(\mathfrak{A})$ in the locally convex topology τ_{loc} induced by the family

$$\{\|\cdot\|_{\alpha,t,\mathrm{loc}}:\alpha\in\mathscr{D}\times\mathscr{H},\ 0<t\in\mathbf{R}_+\}$$

of semi-norms, where

$$\|b\|_{\alpha,t,\mathrm{loc}}^2=\int_0^t ds\|b(s)\|_\alpha^2.$$

Here, b is τ_2-locally square integrable, $\alpha\in\mathscr{D}\times\mathscr{H}$, $0<t\in\mathbf{R}_+$.

Quantum stochastic integrals: Let $f, g\in L_{\mathscr{H}}^{\infty,\mathrm{loc}}(\mathbf{R}_+)$ and $\pi\in L_{\mathscr{B}(\mathscr{H})}^{\infty,\mathrm{loc}}(\mathbf{R}_+)$. For $t\in\mathbf{R}_+$, define $A_f(t)$, $A_g^*(t)$, and $\Lambda_\pi(t)$ as follows:

$$A_f(t)=a(f\chi_{[0,t)})\otimes\mathbf{1}^t,$$

$$A_g^*(t)=a^*(g\chi_{[0,t)})\otimes\mathbf{1}^t,$$

$$\Lambda_\pi(t)=\lambda(\pi\chi_{[0,t)})\otimes\mathbf{1}^t.$$

Then, by Ref. 10, Proposition 4.1, the processes A_f, A_g^*, and Λ_π lie in $\mathrm{Ad}(\mathfrak{A})_{\mathrm{loc}}$ and hence in $L^2(\mathfrak{A},\tau_{\mathrm{loc}})$.

In Ref. 10, Hudson and Parthasarathy develop stochastic integration with respect to the processes A_f, A_g^*, and Λ_π: we refer to Ref. 10 for details. Of interest to us in the sequel is the following result.

Theorem (4.1): (Hudson and Parthasarathy[10]): Let E, F, G, H lie in $L^2(\mathfrak{A},\tau_{\mathrm{loc}})$ and X be the process given by

$$(*)\quad X(t)=\int_0^t\big(E(r)d\Lambda_\pi(r)+F(r)dA_f(r)+G(r)dA_g^*(r)+H(r)dr\big),$$

$t\in\mathbf{R}_+$. Then, for $0\le s\le t<\tau$ and each $\alpha\in\mathscr{D}\times\mathscr{H}$,

$$\|X(t)-X(s)\|_\alpha^2\le c_\alpha(\tau)\int_s^t dr\,e^{t-r}\{\|E(r)\|_\alpha^2+\|F(r)\|_\alpha^2+\|G(r)\|_\alpha^2+\|H(r)\|_\alpha^2\},$$

where $c_\alpha(\tau)$ is a positive constant that depends also on f, g, π. ■

Remark: (1) Notice that, from the above result, it follows that X lies in $\mathrm{Ad}(\mathfrak{A})_{\mathrm{con}}$ and hence also in $L^2(\mathfrak{A},\tau_{\mathrm{loc}})$.

(2) It is for the stochastic integral given by (*) that we present a generalization of the *Ito formula* in the sequel. To this end, we shall utilize the following notion of *adaptedness* of maps.

Definition: We call a map θ: $\mathbf{R}_+\times\mathfrak{A}(\tau_2)\to L_b^{(0)}((\mathfrak{A}(\tau_2))^m,\mathfrak{A}(\tau_2))$ adapted provided that

(1) for each $t\in\mathbf{R}_+$, $\theta(t,Y)(P_1,P_2,\ldots,P_m)$ lies in $\mathfrak{A}_t(\tau_2)$ whenever $Y\in\mathfrak{A}_t(\tau_2)$ and $P_j\in\mathfrak{A}_t$, and $j=1,2,\ldots,m$; and

(2) for each $t\in\mathbf{R}_+$, $Y\in\mathfrak{A}_t(\tau_2)$, $R_j=P_jQ_j\in\mathfrak{A}$ with $P_j\in\mathfrak{A}_t$ and $Q_j\in\mathfrak{A}^t$, $j=1,2,\ldots,m$, we have

$$\theta(t,Y)(R_1,R_2,\ldots,R_m)=\theta(t,Y)(P_1,P_2,\ldots,P_m)\cdot Q_1Q_2\cdots Q_m.$$

Remark: (1) The notion of adaptedness for the map

$$\theta:\mathfrak{A}(\tau_2)\to L_b^{(0)}((\mathfrak{A}(\tau_2))^m,\mathfrak{A}(\tau_2))$$

is analogous to that introduced above.

(2) We now state the main results of this paper.

Theorem (4.2) (Ito formula): Let $\tau>0$ be fixed and $\psi\in C^{1,2}([0,\tau]\times\mathfrak{A}(\tau_2),\mathfrak{A}(\tau_2))$, with partial derivatives $\dot{\psi}$, ψ', and ψ''. Suppose that

(4.2.1) the maps $\dot{\psi}$, ψ', and ψ'' are continuous from $[0,\tau]\times\mathfrak{A}(\tau_2)$ into $\mathfrak{A}(\tau_2)$, $\mathscr{L}_b^{(0)}(\mathfrak{A}(\tau_2))$ and $\mathscr{L}_b^{(0)}((\mathfrak{A}(\tau_2))^2, \mathfrak{A}(\tau_2))$, respectively;

(4.2.2) the maps ψ' and ψ'' are adapted;

(4.2.3) $E, F, G, H\in \mathrm{Ad}(\mathfrak{A})_{\mathrm{con}}$;

(4.2.4) $X_0\in\mathfrak{A}(\tau_2)$ and

$$X(t)=X_0+\int_0^t\{E(s)d\Lambda_\pi(s)+F(s)dA_f(s)+G(s)dA_g^*(s)+H(s)ds\},\quad t\in[0,\tau].$$

Then

$$\begin{aligned}&\psi(t,X(t))-\psi(0,X_0)\\&=\int_0^t ds\{\dot{\psi}(s,X(s))+\psi'(s,X(s))[H(s)]+\psi''(s,X(s))(F(s),G(s))\}\\&\quad+\int_0^t\{\dot{\psi}(s,X(s))[E(s)]d\Lambda_\pi(s)+\psi'(s,X(s))[F(s)]dA_f(s)+\psi'(s,X(s))[G(s)]dA_g^*(s)\}\\&\quad+\int_0^t\{\psi''(s,X(s))(E(s),E(s))d\Lambda_\pi(s)+\psi''(s,X(s))(F(s),E(s))dA_f(s)+\psi''(s,X(s))(E(s),G(s))dA_g^*(s)\},\ t\in[0,\tau].\end{aligned}$$

■

Theorem (4.3): Let $\xi\in C^2(\mathfrak{A}(\tau_2))$ with partial derivatives ξ' and ξ''. Suppose that the maps ξ' and ξ'' are adapted and are continuous from $\mathfrak{A}(\tau_2)$ into $\mathscr{L}_b^{(0)}(\mathfrak{A}(\tau_2))$ and $\mathscr{L}_b^{(0)}((\mathfrak{A}(\tau_2))^2, \mathfrak{A}((\tau_2))$, respectively.

Then, for fixed $\tau>0$, $t\in[0,\tau]$ and all real numbers α,β,γ,

$$\begin{aligned}&\xi(\alpha\Lambda_\pi(t)+\beta A_f(t)+\gamma A_g^*(t))-\xi(0)\\&=\int_0^t\{\beta\gamma\xi''(\alpha\Lambda_\pi(s)+\beta A_f(s)+\gamma A_g^*(s))((1,1)ds+(\alpha\xi'(\alpha\Lambda_\pi(s)+A_f(s)+\gamma A_g^*(s))[1]\\&\quad+\alpha^2\xi''(\alpha\Lambda_\pi(s)+\beta A_f(s)+\gamma A_g^*(s))(1,1))d\Lambda_\pi(s)+(\beta\xi'(\alpha\Lambda_\pi(s)+\beta A_f(s)+\gamma A_g^*(s))[1]\\&\quad+\alpha\beta\xi''(\alpha\Lambda_\pi(s)+\beta A_f(s)+\gamma A_g^*(s))(1,1))dA_f(s)+(\gamma\xi'(\alpha\Lambda_\pi(s)+\beta A_f(s)+\gamma A_g^*(s))[1]\\&\quad+(\alpha\gamma\xi''(\alpha\Lambda_\pi(s)+\beta A_f(s)+\gamma A_g^*(s))(1,1))dA_g^*(s)\}.\end{aligned}$$

■

Notation: Let $\eta\in\mathscr{L}_b^{(0)}((\mathfrak{A}(\tau_2))^m,\mathfrak{A}(\tau_2))$. Then, by the separate continuity of η and Ref. 23, Theorem III.1.1, it follows that to each $\alpha\in\mathscr{D}\times\mathscr{H}$, there corresponds a positive number $c_{\eta,\alpha}$, depending on η, and members $\mathbf{H}_1(\alpha),\ldots,\mathbf{H}_m(\alpha)\in\mathrm{Fin}(\mathscr{D}\times\mathscr{H})$, independent of η, such that

$$\|\eta(x_1,\ldots,x_m)\|_\alpha\leqslant c_{\eta,\alpha}\|x_1\|_{\mathbf{H}_1(\alpha)}\cdots\|x_m\|_{\mathbf{H}_m(\alpha)},\qquad(4.4)$$

for all $(x_1,\ldots,x_m)\in[\mathfrak{A}(\tau_2)]^m$.

Now, let $\Theta_1,\Theta_2,\ldots,\Theta_m$ be bounded subsets of $\mathfrak{A}(\tau_2)$. Then, we use the notation:

$$|||\eta|||_{\alpha,\Theta_1,\ldots,\Theta_m}=\inf\{c_{\eta,\alpha}:\ (4.4)\ \text{holds}\ \forall x_j\in\Theta_{j,j=1,2,\ldots,m}\}.\qquad(4.5)$$

Remark: In what follows, we prove only Theorem (4.2) because the proof of Theorem (4.3) is only slightly different, in view of the boundedness in $\mathfrak{A}(\tau_2)$ of the set:

$$\{\Lambda_\pi(t_2)-\Lambda_\pi(t_1),A_f(t_4)-A_f(t_3),A_g^*(t_6)-A_g^*(t_5):\ 0\leqslant t_1\leqslant t_2\leqslant\tau,\ 0\leqslant t_3\leqslant t_4\leqslant\tau,\ 0\leqslant t_5\leqslant t_6\leqslant\tau\}.$$

Proposition 4.6: Let the hypotheses of Theorem (4.2) hold with E, F, G, H lying in $\mathrm{Ad}(\mathfrak{A})_{\mathrm{sim}}$. Suppose that $0=t_0<t_1<\cdots<t_N=t$ is a partition of $[0,t]$, $t\in(0,\tau]$. Then, the sum

$$(4.6)_{N,t}\quad\sum_{n=1}^N\psi'(t_{n-1},X(t_{n-1}))[X(t_n)-X(t_{n-1})]$$

converges in $\mathfrak{A}(\tau_2)$ to the expression

$$\begin{aligned}(4.6)_{\infty,t}\quad&\int_0^t\{\psi'(s,X(s))[H(s)]ds\\&+\psi'(s,X(s))[E(s)]d\Lambda_\pi(s)\\&+\psi'(s,X(s))[F(s)]dA_f(s)\\&+\psi'(s,X(s))[G(s)]dA_g^*(s)\},\quad 0<t\leqslant\tau.\end{aligned}$$

Proof: Let $0=t_0<t_1<\cdots<t_N=t$ be a partition of $[0,t]$, $0<t\leqslant\tau$, and

$$E(\cdot)=\sum_{n=0}^\infty E_n\chi_{[t_n,t_{n+1})}(\cdot),$$

$$F(\cdot)=\sum_{n=0}^\infty F_n\chi_{[t_n,t_{n+1})}(\cdot),$$

$$G(\cdot)=\sum_{n=0}^\infty G_n\chi_{[t_n,t_{n+1})}(\cdot),$$

$$H(\cdot)=\sum_{n=0}^\infty H_n\chi_{[t_n,t_{n+1})}(\cdot),$$

with $E_n,F_n,G_n,H_n\in\mathfrak{A}_{t_n}$, $n=0,1,2,\ldots$. Then

$$\sum_{n=1}^{N} \psi'(t_{n-1},X(t_{n-1}))[X(t_n)-X(t_{n-1})]$$

$$= \sum_{n=1}^{N} \psi'(t_{n-1},X(t_{n-1}))\left[\int_{t_{n-1}}^{t_n} \{H(s)ds + E(s)d\Lambda_\pi(s) + F(s)dA_f(s) + G(s)dA_g^*(s)\right]$$

$$= \sum_{n=1}^{N} \{\psi'(t_{n-1},X(t_{n-1}))[H(t_{n-1})](t_n - t_{n-1}) + \psi'(t_{n-1},X(t_{n-1}))[E(t_{n-1})](\Lambda_\pi(t_n) - \Lambda_\pi(t_{n-1}))$$

$$+ \psi'(t_{n-1},X(t_{n-1}))[F(t_{n-1})](A_f(t_n) - A_f(t_{n-1})) + \psi'(t_{n-1},X(t_{n-1}))[G(t_{n-1})](A_g^*(t_n) - A_g^*(t_{n-1}))\},$$

by the adaptedness of ψ'. Then, using once more the adaptedness of the maps ψ' and the definition of the stochastic integral, we see that $(4.6)_{N,t}$ converges in $\mathfrak{A}(\tau_2)$ to $(4.6)_{\infty,t}$, for each $t\in[0,t]$.

■

Proposition 4.7: Let the hypotheses of Theorem (4.2) hold. Then as $N\to\infty$, with $\max_{1<n<N}|t_n - t_{n-1}|\to 0$, the expression $(4.6)_{N,t}$ converges in $\mathfrak{A}(\tau_2)$ to the expression $(4.6)_{\infty,t}$.

Proof: By Ref. 10, Proposition 3.2, every member of $L^2(\mathfrak{A},\tau_{\rm loc})$ is the $\tau_{\rm loc}$-limit of a sequence of members of $\mathrm{Ad}(\mathfrak{A})_{\rm sim}$. So, let $\{E_m\}_{1<m<\infty}$, $\{F_m\}_{1<m<\infty}$, $\{G_m\}_{1<m<\infty}$, and $\{H_m\}_{1<m<\infty}$ be sequences in $\mathrm{Ad}(\mathfrak{A})_{\rm sim}$ which converge in $L^2(\mathfrak{A},\tau_{\rm loc})$ to E, F, G, and H, respectively. Since $\tau_{\rm loc}$ is finer than $\tau_{\rm loc}$ and E, F, G, H are τ_2-continuous, by hypothesis, and hence are contained in $\mathrm{Ad}(\mathfrak{A},\tau_\infty)$, we may assume that $\{E_m\}_{1<m<\infty}$, $\{F_m\}_{1<m<\infty}$, $\{G_m\}_{1<m<\infty}$, and $\{H_m\}_{1<m<\infty}$ are contained in $\mathrm{Ad}(\mathfrak{A},\tau_\infty)$ and converge there to E, F, G, and H, respectively. Then, $\{X_m\}_{1<m<\infty}$ is contained in $\mathrm{Ad}(\mathfrak{A},\tau_\infty)$ and converges there to X, where

$$X_m(t) = \int_0^t \{H_m(s)ds + E_m(s)d\Lambda_\pi(s) + F_m(s)dA_f(s) + G_m(s)dA_g^*(s)\}$$

and

$$X(t) = \int_0^t \{H(s)ds + E(s)d\Lambda_\pi(s) + F(s)A_f(s) + G(s)dA_g^*(s)\},\ t\ [0,\tau].$$

Hence, for some sufficiently large m_0, there is a bounded subset Θ_{m_0} of $\mathfrak{A}(\tau_2)$ such that

(4.7.1) $\{E(s),F(s),G(s),H(s),X(s):0\leqslant s<\tau\}\cup\{E_m(s),F_m(s),G_m(s),H_m(s):0\leqslant s\leqslant\tau\}\cup$

$\cup\{E_m(s)-E(s),F_m(s)-F(s),G_m(s)-G(s),H_m(s)-H(s):0\leqslant s\leqslant\tau\}\cup$

$\cup\{X_m(s)-X(s):0\leqslant s\leqslant\tau\}\subset\Theta_{m_0}$, for all $m\geqslant m_0$.

Set

(4.7.2) $$\sup_{0<s<\tau}\sup_{x\in\Theta_{m_0}} |||\psi'(s,x)|||_{\alpha,\Theta_{m_0}} = C_{\alpha,m_0}\alpha\in\mathscr{D}\times\mathscr{H}.$$

Then, C_{α,m_0} is finite for each α, since $(t,x)\to\psi'(t,x)$ is continuous from $[0,\tau]\times\mathfrak{A}(\tau_2)$ into $\mathscr{L}_b^{(0)}(\mathfrak{A}(\tau_2))$.

By Proposition (4.6), the expression of $(4.6)_{\infty,t}$ is valid when the integrands lie in $\mathrm{Ad}(\mathfrak{A})_{\rm sim}$. Now, let $(4.6)_{\infty,t}^{(m)}$ be the expression $(4.6)_{\infty,t}$ with E, F, G, H, X replaced by E_m,F_m,G_m,H_m,X_m, respectively. Then, we shall show that $(4.6)_{\infty,t}^{(m)}$ converges in $\mathfrak{A}(\tau_2)$ to $(4.6)_{\infty,t}$ as $m\to\infty$. To this end for each $\alpha\in\mathscr{D}\times\mathscr{H}$, we have

$$\|(4.6)_{\infty,t}^{(m)} - (4.6)_{\infty,t}\|_\alpha^2 = \| \int_0^t \{(\psi'(s,X_m(s))[H_m(s)] - \psi'(s,X(s))[H(s)])ds + (\psi'(s,X_m(s))[E_m(s)]$$

$$- \psi'(s,X(s))[E(s)])d\Lambda_\pi(s) + (\psi'(s,X_m(s))[F_m(s)] - \psi'(s,X(s)[F(s)])dA_f(s)$$

$$+ (\psi'(s,X_m(s))[G_m(s)] - \psi'(s,X(s))[G(s)])dA_g^*(s)\}\|_\alpha^2$$

$$\leqslant 2c_\alpha(\tau)e^\tau \int_0^t ds(|||\psi'(s,X_m(s))|||_{\alpha,\Theta_{m_0}}^2\{\|H_m(s)-H(s)\|_{\mathbf{H}(\alpha)}^2 + \|E_m(s)-E(s)\|_{\mathbf{H}(\alpha)}^2$$

$$+ \|F_m(s)-F(s)\|_{\mathbf{H}(\alpha)}^2 + \|G_m(s)-G(s)\|_{\mathbf{H}(\alpha)}^2\} + |||\psi'(s,X_m(s)) - \psi'(s,X(s))|||_{\alpha,\Theta_{m_0}}^2\{\|H(s)\|_{\mathbf{H}(\alpha)}^2$$

$$+ \|E(s)\|_{\mathbf{H}(\alpha)}^2 + \|F(s)\|_{\mathbf{H}(\alpha)}^2 + \|G(s)\|_{\mathbf{H}(\alpha)}^2\})$$

where $\mathbf{H}(\alpha)$ is some member of $\mathrm{Fin}(\mathscr{D}\times\mathscr{H})$ depending only on α, and using (4.1) and (4.5),

$$\leqslant 2c_\alpha(\tau)e^\tau C_{\alpha,m_0} \int_0^t ds(\|H_m(s)-H(s)\|_{\mathbf{H}(\alpha)}^2 + \|E_m(s)-E(s)\|_{\mathbf{H}(\alpha)}^2 + \|F_m(s)-F(s)\|_{\mathbf{H}(\alpha)}^2 + \|G_m(s)-G(s)\|_{\mathbf{H}(\alpha)}^2)$$

$$+ 2c_\alpha(\tau)e^\tau \int_0^t ds|||\psi'(s,X_m(s)) - \psi'(s,X(s))|||_{\alpha,\Theta_{m_0}}^2 (\|H(s)\|_{\mathbf{H}(\alpha)}^2 + \|E(s)\|_{\mathbf{H}_{(\alpha)}}^2 + \|F(s)\|_{\mathbf{H}(\alpha)}^2 + \|G(s)\|_{\mathbf{H}(\alpha)}^2),$$

for all $m\geqslant m_0$, where Θ_{m_0} is as described in (4.7.1) and using (4.4), (4.5), and (4.7.2). Thus, using the fact that $\{E_m\}_{1<m<\infty}$, $\{F_m\}_{1<m<\infty}$, $\{G_m\}_{1<m<\infty}$, and $\{H_m\}_{1<m<\infty}$ converge in $\mathrm{Ad}(\mathfrak{A},\tau_\infty)$ and in $L^2(\mathfrak{A},\tau_{\rm loc})$ to the same limits E, F, G, H, respectively, and the continuity of the map ψ': $[0,\tau]\times\mathfrak{A}(\tau_2)\to(\mathfrak{A}(\tau_2))$, we conclude that $(4.6)_{\infty,t}^{(m)}$ converges in $\mathfrak{A}(\tau_2)$ to $(4.6)_{\infty,t}$ for each $t\in[0,\tau]$ as $m\to\infty$. This ends the proof.

Proposition (4.8): Let the hypotheses of Theorem (4.2) hold. Suppose that $0 = t_0 < t_1 < \cdots < t_N = t$ is a partition of $[0,t]$, $t \in (0,\tau]$. Then, as $N \to \infty$, with $\max_{1<n<N} |t_n - t_{n-1}| \to 0$, the sum

$$(4.8)_{N,t} := \sum_{n=1}^{N} \psi''(t_{n-1}, X(t_{n-1}))(X(t_n) - X(t_{n-1}), X(t_n) - X(t_{n-1}))$$

converges in $\mathfrak{A}(\tau_2)$ to the expression

$$(4.8)_{\infty,t} := \int_0^t \{\psi''(s,X(s))(F(s),G(s))ds + \psi''(s,X(s))(E(s),E(s))d\Lambda_\pi(s)$$
$$+ \psi''(s,X(s))(F(s),E(s))dA_f(s) + \psi''(s,X(s))(E(s),G(s))dA_g^*(s)\}, \quad t \in (0,t].$$

Proof: As in the proof of Proposition (4.6), it is easy to show that the assertion of the present Proposition is valid when E, F, G, H all lie in $\mathrm{Ad}(\mathfrak{A})_{\mathrm{sim}}$.

To complete the proof, let $\{E_m\}_{1<m<\infty}$, $\{F_m\}_{1<m<\infty}$, $\{G_m\}_{1<m<\infty}$, $\{H_m\}_{1<m<\infty}$, and $\{X_m\}_{1<m<\infty}$ be as in the proof of Proposition (4.7), and write $(4.8)_{\infty,t}^{(m)}$ for the expression $(4.8)_{\infty,t}$ with E, F, G, H, X replaced by E_m, F_m, G_m, H_m, X_m, respectively. Then, we shall show that $(4.8)_{\infty,t}^{(m)}$ converges in $\mathfrak{A}(\tau_2)$ to $(4.8)_{\infty,t}$, as $m \to \infty$. To this end, for each $\alpha \in \mathscr{D} \times \mathscr{H}$, we have

$$\|(4.8)_{\infty,t}^{(m)} - (4.8)_{\infty,t}\|_\alpha^2$$
$$= \left|\left| \int_0^t \{\psi''(s,X_m(s))(E_m(s),E_m(s)) - \psi''(s,X(s))(E(s),E(s))\}d\Lambda_\pi(s) \right.\right.$$
$$+ (\psi''(s,X_m(s))(F_m(s),E_m(s)) - \psi''(s,X(s))(F(s),E(s)))dA_f(s)$$
$$+ (\psi''(s,X_m(s))(E_m(s),G_m(s)) - \psi''(s,X(s))(E(s),G(s)))dA_g^*(s)$$
$$\left.\left. + (\psi''(s,X_m(s))(F_m(s),G_m(s)) - \psi''(s,X(s))(F(s),G(s)))ds\} \right|\right|_\alpha^2$$
$$\leqslant e^\tau c_\alpha(\tau) \int_0^t ds\{\|\psi''(s,X_m(s))(E_m(s),E_m(s)) - \psi''(s,X(s))(E(s),E(s))\|_\alpha^2$$
$$+ \|\psi''(s,X_m(s))(F_m(s),E_m(s)) - \psi''(s,X(s))(F(s),E(s))\|_\alpha^2$$
$$+ \|\psi''(s,X_m(s))(E_m(s),G_m(s)) - \psi''(s,X(s))(E(s),G(s))\|_\alpha^2$$
$$+ \|\psi''(s,X_m(s))(F_m(s),G_m(s)) - \psi''(s,X(s))(F(s),G(s))\|_\alpha^2\},$$

using Theorem (4.1) and the adaptedness of ψ''. The above integral may readily be estimated. To indicate the trend of argument, let us estimate the integral of the first integrand.

Set

$$\sup_{c<s<\tau} \sup_{x \in \Theta_{m_0}} |||\psi''(s,x)|||_{\alpha,\Theta_{m_0},\Theta_{m_0}} = \lambda_{\alpha,m_0}.$$

Then, $\lambda_{\alpha,m_0} < \infty$, since the map $\psi'':[0,\tau] \times \mathfrak{A}(\tau_2) \to \mathscr{L}_b^{(0)}((\mathfrak{A}(\tau_2))^2, \mathfrak{A}(\tau_2))$ is continuous.

Now

$$(4.8)_{\infty,t,E}^{(m)} \equiv \int_0^t ds\|\psi''(s,X_m(s))(E_m(s)E_m(s)) - \psi''(s,X(s))(E(s),E(s))\|_\alpha^2$$
$$\leqslant \int_0^\infty ds\{|||\psi''(s,X_m(s)) - \psi''(s,X(s))|||_{\alpha,\Theta_{m_0},\Theta_{m_0}}^2 \cdot \|E(s)\|_{\mathbf{H}_1(\alpha)}^2 \|E(s)\|_{\mathbf{H}_2(\alpha)}^2$$
$$+ |||\psi''(s,X_m(s))|||_{\alpha,\Theta_{m_0},\Theta_{m_0}}^2 (\|E(s) - E_m(s)\|_{\mathbf{H}_1(\alpha)}^2 \|E(s) - E_m(s)\|_{\mathbf{H}_2(\alpha)}^2$$
$$+ \|E(s)\|_{\mathbf{H}_1(\alpha)}^2 \|E(s) - E_m(s)\|_{\mathbf{H}_2(\alpha)}^2 + \|E(s) - E_m(s)\|_{\mathbf{H}_1(\alpha)}^2 \|E(s)\|_{\mathbf{H}_2(\alpha)}^2)\}$$
$$\leqslant \left(\sup_{0<s<\tau} \|E(s)\|_{\mathbf{H}_1(\alpha)}^2\right) \int_0^t ds|||\psi''(s,X_m(s)) - \psi''(s,X(s))|||_{\alpha,\Theta_{m_0},\Theta_{m_0}}^2 \|E(s)\|_{\mathbf{H}_2(\alpha)}^2$$
$$+ \lambda_{\alpha,m_0}^2 \{\sup_{0<s<\tau} \|E(s) - E_m(s)\|_{\mathbf{H}_1(\alpha)}^2 \int_0^t ds\|E(s) - \|E_m(s)\|_{\mathbf{H}_2(\alpha)}^2$$
$$+ \sup_{0<s<t} \|E(s)\|_{\mathbf{H}_1(\alpha)}^2 \int_0^t ds\|E(s) - E_m(s)\|_{\mathbf{H}_2(\alpha)}^2 + \sup_{0<s<t} \|E(s)\|_{\mathbf{H}_2(\alpha)}^2 \int_0^t ds\|E(s) - E_m(s)\|_{\mathbf{H}_1(\alpha)}^2\},$$

for a sufficiently large m_0 and all $m \geqslant m_0$, with Θ_{m_0} as in Proposition (4.7). Using the fact that $\{E_m\}_{1<m<\infty}$ converges both in $\mathrm{Ad}(\mathfrak{A},\tau_\infty)$ and in $L^2(\mathfrak{A},\tau_{\mathrm{loc}})$ to the same limit E, we conclude that $(4.8)_{\infty,t,E}^{(m)}$ converges to zero as $m \to \infty$, for each $t \in (0,\tau]$.

The other integrals in the estimate for $\|(4.8)^{(m)}_{\infty,t} - (4.8)_{\infty,t}\|^2_\alpha$ are similarly shown to converge to zero for each $t\in(0,\tau]$ as $m_0\to\infty$. Hence $(4.8)^{(m)}_{\infty,t}$ converges in $\mathfrak{A}(\tau_2)$ to $(4.8)_{\infty,t}$ for each $t\in(0,t]$. This concludes the proof. ■

Proof of Theorem (4.2): Let $0=t_0<t_1<\cdots<t_N=t$ be a partition of $[0,t]$, $t\in(0,\tau]$. Then

$$
\begin{aligned}
&\psi(t,X(t))-\psi(0,X_0)\\
&\quad=\sum_{n=1}^{N}(\psi(t_n,X(t_n))-\psi(t_{n-1},X(t_{n-1})))\\
&\quad=\sum_{n=1}^{N}(\psi(t_n,X(t_n))-\psi(t_{n-1},X(t_n)))+\sum_{n=1}^{N}(\psi(t_{n-1},X(t_n))-\psi(t_{n-1},X(t_{n-1})))\\
&\quad=\sum_{n=1}^{N}(t_n-t_{n-1})\dot\psi(t_{n-1},X(t_n))+\sum_{n=1}^{N}\psi'(t_{n-1},X(t_{n-1}))[X(t_n)-X(t_{n-1})]\\
&\qquad+\frac{1}{2}\sum_{n=1}^{N}\psi''(t_{n-1},X(t_{n-1}))(X(t_n)-X(t_{n-1}),X(t_n)-X(t_{n-1}))\\
&\qquad+\sum_{n=1}^{N}(t_n-t_{n-1})\int_0^1 d\alpha\{\dot\psi(t_{n-1}+(t_n-t_{n-1})\alpha,X(t_n))-\dot\psi(t_{n-1},X(t_n))\}\\
&\qquad+\sum_{n=1}^{N}\int_0^1 d\alpha(1-\alpha)\Gamma(t_{n-1},t_n,X,\alpha)(X(t_n)-X(t_{n-1}),X(t_n)-X(t_{n-1})),
\end{aligned}
$$

where

$$\Gamma(t_{n-1},t_nX,\alpha)=\psi''(t_{n-1},X(t_{n-1})+(X(t_n)-X(t_{n-1}))\alpha)-\psi''(t_{n-1},X(t_{n-1}))$$

Now, in view of (4.2.1), the sum $\Sigma_{n=1}^{N}(t_n-t_{n-1})\dot\psi(t_{n-1},X(t_n))$ converges in $\mathfrak{A}(\tau_2)$ to $\int_0^t ds\,\dot\psi(s,X(s))$ as $N\to\infty$, with $\max_{1\le n\le N}|t_n-t_{n-1}|\to 0$.

Next, for each $\alpha\in\mathscr{D}\times\mathscr{H}$, we have

$$
\begin{aligned}
&\left|\left|\sum_{n=1}^{N}(t_n-t_{n-1})\int_0^1 d\alpha\{\dot\psi(t_{n-1}+(t_n-t_{n-1})\alpha,X(t_n))-\dot\psi(t_{n-1},X(t_n))\}\right|\right|_\alpha\\
&\quad\le\Big(\max_{1\le n\le N}|t_n-t_{n-1}|\Big)\sum_{n=1}^{N}\sup_{0<\alpha<1}\|\dot\psi(t_{n-1}+(t_n-t_{n-1})\alpha,X(t_n)-\dot\psi(t_{n-1},X(t_n))\|_\alpha
\end{aligned}
$$

and

$$
\begin{aligned}
&\left|\left|\sum_{n=1}^{N}\int_0^1 d\alpha(1-\alpha)\Gamma(t_{n-1},t_n,X,\alpha)(X(t_n)-X(t_{n-1}),X(t_n)-X(t_{n-1}))\right|\right|_\alpha\\
&\quad\le\Big(\max_{1\le n\le N}\|X(t_n)-X(t_{n-1}\|_{\mathbf{H}_1(\alpha)}\Big)\cdot\Big(\max_{1\le n\le N}\|X(t_n)-X(t_{n-1})\|_{\mathbf{H}_2(\alpha)}\Big)\sum_{n=1}^{N}\sup_{0<\alpha<1}\|\Gamma(t_{n-1},t_n,X,\alpha)\|_{\alpha,\Theta_X,\Theta_X},
\end{aligned}
$$

where Θ_X is the subset

$$\Theta_X=\{X(t)-X(s):0\le s<t\le\tau\}$$

of $\mathfrak{A}(\tau_2)$, which, by Theorem (4.1), is bounded in $\mathfrak{A}(\tau_2)$, and using the notation of (4.5).

Hence, using (4.2.1) and the τ_2-continuity of X, it follows that as $N\to\infty$, with $\max_{1\le n\le N}|t_n-t_{n-1}|\to 0$, the sums

$$
\begin{aligned}
&\sum_{n=1}^{N}(t_n-t_{n-1})\int_0^1 d\alpha\{\dot\psi(t_{n-1}+(t_n-t_{n-1})\alpha,\\
&\quad X(t_n))-\dot\psi(t_{n-1},X(t_n))\}
\end{aligned}
$$

and

$$
\begin{aligned}
&\sum_{n=1}^{N}\int_0^1 d\alpha(1-\alpha)\Gamma(t_{n-1},t_n,X,\alpha)(X(t_n)-X(t_{n-1}),\\
&\quad X(t_n)-X(t_{n-1}))
\end{aligned}
$$

converge in $\mathfrak{A}(\tau_2)$ to zero.

Finally, by applying Propositions (4.7) and (4.8) to the second and third sums in the expression for $\psi(t,X(t))-\psi(0,X_0)$, the assertion of Theorem (4.2) is obtained. This concludes the proof. ■

Remark: (1) Let $\mathscr{H}^n$ be the n-fold direct sum of $\mathscr{H}$ with itself, $\mathbf{X}=\mathbf{R}\otimes\Gamma(\mathscr{H}^n)$, where $\mathbf{R}$ is some fixed Hilbert space, and $\Omega(\mathbf{f})$ be an exponential vector in $\Gamma(\mathscr{H}^n)$, for $\mathbf{f}\in\mathscr{H}^n$. For $f_j\in\mathscr{H}$ and $\pi_j\in L^{\infty,\mathrm{loc}}_{B(\mathscr{H})}(\mathbb{R}_+)$, define the operators $a_j(f_j)$, $a_j^*(f_j)$, and $\lambda_j(\pi_j)$, $j=1,2,\ldots,n$, on the linear span $V\{\Omega(\mathbf{f}):\mathbf{f}\in\mathscr{H}^n\}$ as follows:

$$a_j(f_j)\Omega(\mathbf{g})=\langle f_j,g_j\rangle_{\mathscr{H}}\Omega(\mathbf{g})$$

$$a_j^*(f_j)\Omega(\mathbf{g})=\frac{d}{dx}\Omega(\mathbf{g}+\overset{n}{\underset{k=1}{\oplus}}\delta_{kj}xf_k)|_{x=0}$$

$$\lambda_j(\pi_j)\Omega(\mathbf{g})=\frac{d}{dx}\Omega(g_1\oplus\cdots\oplus g_{j-1}\oplus e^{x\pi_j}g_j\oplus\cdots\oplus g_n)|_{x=0}$$

$$\mathbf{g}=\overset{n}{\underset{j=1}{\oplus}}g_j\in\mathscr{H}^n,\quad j=1,2,\ldots,n.$$

In a manner analogous to how $\mathfrak{S}$ was constructed in Sec. II, we can construct a dense subspace $\mathfrak{S}$, containing $V\{\Omega(\mathbf{f}):\mathbf{f}\in\mathscr{H}^n\}$, which is invariant under the action of any member of the polynomial algebra generated by

$$\{a_j(f_j),a_j^*(g_j),\lambda_j\pi_j:f_j,g_j\in\mathscr{H},\pi_j\in L^{\infty,\mathrm{loc}}_{B(\mathscr{H})}(\mathbf{R}_+),\ j=1,2,\ldots,n\}.$$

Let $\mathbf{D}$ be a fixed dense subspace of $\mathscr{R}$. Then $\mathbf{E}\equiv\mathbf{D}\otimes\mathfrak{S}$ is dense in $\mathscr{H}^n$. We set $\mathfrak{A}(\mathbf{E})\equiv\mathfrak{A}$, $\mathfrak{A}(\mathbf{E}_t)\otimes\mathbb{1}^t\equiv\mathfrak{A}_t$ and $\mathbb{1}_t\otimes\mathfrak{A}(\mathfrak{S}^t)=\mathfrak{A}^t$ and denote the τ_2-completions of these by $\mathfrak{A}(\tau_2)$, $\mathfrak{A}_t(\tau_2)$, and $\mathfrak{A}^t(\tau_2)$, respectively. For f_j,g_k and $\pi_j\in L^{\infty,\mathrm{loc}}_{B(\mathscr{H})}(\mathbf{R}_+)$, define $A_{f_j}(t)$, $A^*_{f_j}(t)$, and $\Lambda_{\pi_j}(t)$, $t\in\mathbf{R}_+$, by

$$A_{f_j}(t)=a_j(f_j\chi_{[0,t)})\otimes\mathbb{1}^t,\quad A^*_{g_j}(t)=a_j^*(g_j\chi_{[0,t)})\otimes\mathbb{1}^t,$$

$$\Lambda_{\pi_j}(t)=\lambda_j(\pi_j\chi_{[0,t)})\otimes\mathbb{1}^t,$$

$j=1,2,\ldots,n$. These operators feature in Theorem (4.9) below.

(2) If $\psi\in C^2((\mathfrak{A}(\tau_2))^n,\mathfrak{A}(\tau_2))$ and $Y,P,Q\in(\mathfrak{A}(\tau_2))^n$, with $P=(P_1,\ldots,P_n)$ and $Q=(Q_1,\ldots,Q_n)$, then $\psi'(Y)[P]$ and $\psi''(Y)(P,Q)$ may be written as follows:

$$\psi'(Y)[p]=\Sigma_{j=1}^n\psi_{,j}(Y)[P_j],$$

where $\psi'_{,j}(Y)\in\mathscr{L}_b^{(0)}(\mathfrak{A}(\tau_2))$ is the jth partial derivative of ψ at the point Y, and

$$\psi''(Y)(P,Q)=\Sigma_{j=1}^n\Sigma_{k=1}^n\psi''_{jk}(Y)(P_k,Q_j),$$

where $\psi''_{jk}(Y)\in\mathscr{L}_b^{(0)}((\mathfrak{A}(\tau_2))^2,\mathfrak{A}(\tau_2))$ is the (j,k)-mixed partial derivative of ψ at the point Y.

(3) The d-dimensional form of the Ito formula is established by methods analogous to those already indicated above in the case $d=1$. Since the requisite notation becomes considerably more involved, we shall only state the result.

Theorem (4.9): Let $\tau>0$ be fixed and $\psi\in C^{1,2}([0,\tau]\times(\mathfrak{A}(\tau_2))^n,\mathfrak{A}(\tau_2))$, with partial derivatives $\dot\psi$, ψ'_j, ψ''_{jk}, $1\leqslant j,k\leqslant n$. Suppose that

(i) the maps $\dot\psi$, ψ'_j, and ψ''_{jk}, $1\leqslant j,k\leqslant$ are continuous from $[0,\tau]\times(\mathfrak{A}(\tau_2))^n$ into $\mathfrak{A}(\tau_2)$, $\mathscr{L}_b^{(0)}(\mathfrak{A}(\tau_2))$, and $\mathscr{L}_b^{(0)}((\mathfrak{A}(\tau_2))^2,\mathfrak{A}(\tau_2))$, respectively;

(ii) the maps ψ'_j and ψ''_{jk}, $1\leqslant j,k\leqslant n$, are adapted;

(iii) $\{H_j,E_{jk},F_{jk},G_{jk}:1\leqslant j\leqslant d,1\leqslant k\leqslant n\}\subset\mathrm{Ad}(\mathfrak{A})_{\mathrm{con}}$;

(iv) $\mathbf{X}_0=(X_{0j})$, with $X_{0j}\in\mathfrak{A}(\tau_2)$, $1\leqslant j\leqslant d$, and $\mathbf{X}=(X_j)$ with

$$X_j(t)=X_{0j}+\int_0^t\Big\{H_j(s)ds+\sum_{k=1}^n\big(E_{jk}(s)d\Lambda_{\pi_k}(s)+F_{jk}(s)dA_{f_k}(s)+G_{jk}(s)dA^*_{g_k}(s)\big)\Big\},$$

$t\in[0,\tau]$, $1\leqslant j\leqslant d$.

Then

$$\begin{aligned}
&\psi(t,\mathbf{X}(t))-\psi(0,\mathbf{X}_0)\\
&=\int_0^t ds\Big\{\dot\psi(s,X(s))+\sum_{j=1}^d\psi'_j(s,X(s))[H_j(s)]+\sum_{j=1}^d\sum_{l=1}^d\sum_{k=1}^n\psi''_{jl}(s,X(s))(F_{lk}(s),G_{jk}(s))\Big\}\\
&\quad+\sum_{j=1}^d\sum_{k=1}^n\int_0^t\{\psi'_j(s,X(s))[E_{jk}(s)]d\Lambda_{\pi_k}(s)+\psi'_j(s,X(s))[F_{jk}(s)]dA_{f_k}(s)+\psi'_j(s,X(s))[G_{jk}(s)]dA^*_{g_k}(s)\}\\
&\quad+\sum_{j=1}^d\sum_{l=1}^d\sum_{k=1}^d\int_0^t\{\psi''_{jl}(s,X(s))(E_{lk}(s),E_{jk}(s))d\Lambda_{\pi_k}(s)+\psi''_{jl}(s,X(s))(F_{lk}(s),E_{jk}(s))dA_{f_k}(s)\\
&\quad+\psi''_{jl}(s,X(s))(E_{lk}(s),G_{jk}(s))dA^*_{g_k}(s)\},\quad t\in(0,\tau].
\end{aligned}$$

■

Remark: The following immediate consequence of Theorem (4.9) is important in its own right.

Corollary (4.10): Let $\tau>0$ be fixed. Suppose that

(i) $\{H_j,E_{jk},F_{jk},G_{jk}:j=1,2,1\leqslant k\leqslant n\}\in\mathrm{Ad}\mathfrak{A}_{\mathrm{con}}$,

(ii) $\mathbf{X}_0=(X_{0j})$, with $X_{0j}\in\mathfrak{A}$, $j=1,2$, and $\mathbf{X}=(X_j)$, with

$$X_j(t)=X_{0j}+\int_0^t\{H_j(s)ds+\sum_{k=1}^n\big(E_{jk}(s)d\Lambda_{\pi_k}(s)+F_{jk}(s)dA_{f_k}(s)+G_{jk}(s)dA^*_{g_k}(s)\big)\},\quad j=1,2,$$

$t\in(0,\tau]$.

Then

$$\begin{aligned}
X_1(t)X_2(t)&=X_{01}X_{02}+[X_1,X_2]_t\\
&\quad+\int_0^t\{dX_1(s)X_2(s)+X_1(s)dX_2(s)\},
\end{aligned}$$

where

$$[X_1,X_2]_t=\sum_{k=1}^n\int_0^t ds\{F_{1k}(s)G_{2k}(s)+F_{2k}(s)G_{1k}(s)\},\quad t\in(0,\tau].$$

Remark: (1) Notice that $X_j(t)$ lies in $\mathfrak{A}$ for each $t\in\mathbf{R}_+$ and $j=1,2$. Hence the product $X_1(t)X_2(t)$ exists for each $t\in\mathbf{R}_+$.

(2) The above result is an *integration by parts* formula.

(3) The noncommutative extension of the Ito formula presented in Ref. 10, Theorem 4.5 is a special case of the above Corollary since, unlike in Ref. 10, Theorem 4.5, we do not assume that the integrands in the integrals defining X_j, $j=1,2$, are bounded.

(4) Applications of the results of this paper will appear elsewhere.

ACKNOWLEDGMENTS

This work was begun at the Forschungszentrum Bielefeld. I am grateful to my hosts Professor S. Albeverio, Pro-

fessor Ph. Blanchard, and Professor L. Streit for their kind hospitality. I also wish to acknowledge the helpful comments/suggestions of the referee.

Work supported by a Research Fellowship of the Alexander von Humboldt-Stiftung, Federal Republic of Germany.

[1]K. Ito, Proc. Imp. Acad. Tokyo **20**, 519 (1944).
[2]H. Kunita and S. Watanabe, Nagoya Math. J. **30**, 209 (1967).
[3]P. A. Meyer, Illinois J. Math. **6** 193 (1962).
[4]P. A. Meyer, Illinois J. Math. **7** (1963).
[5]N. Ikeda and S. Watanabe, *Stochastic Differential Equations and Diffusion Processes* (North-Holland, Amsterdam, 1981).
[6]K. Ito, Mem. Am. Math. Soc., No. 4 (1961).
[7]D. W. Stroock and S. R. S. Varadhan, Pure Appl. Math. **22**, 345, 479 (1969).
[8]C. Barnett, R. F. Streater, and I. F. Wilde, Funct. Anal. **48**, 172 (1982).
[9]G. O. S. Ekhaguere, Nigerian Math. Soc. **1**, 11 (1982).
[10]R. L. Hudson and K. R. Parthasarathy, Commun. Math. Phys. **93**, 301 (1984).
[11]K. Ito, Nagoya Math. J **3**, 55 (1951).
[12]G. Lassner, Rep. Math. Phys. **3**, 279 (1972).
[13]R. F. Curtain and P. L. Falb, J. Diff. Eqs. **10**, 412 (1971).
[14]B. Gravereaux and J. Pellaumail, Ann. Inst. H. Poincare **10** (4), 339 (1974).
[15]H. H. Kuo, Pacific J. Math. **41**, 469 (1972).
[16]H. K. Kuo, Trans. Am. Math. Soc. **169**, 439 (1972).
[17]D. Kannan and A. T. Bharucha-Reid, Proc. Jpn. Acad. **47**, 472 (1971).
[18]D. Kannan, Ann. Inst. H. Poincare **8**, 9 (1972).
[19]R. L. Hudson and R. F. Streater, Phys. Lett. A **86**, 277 (1981).
[20]H. Hasegawa and R. F. Streater, J. Phys. A: Math. Gen. **16**, L697 (1983).
[21]J. Cook, Trans. Am. Math. Soc. **74**, 222 (1953).
[22]A. Guichardet, *Symmetric Hilbert Spaces and Related Topics*, Lecture Notes in Mathematics Vol. 261 (Springer-Verlag, Berlin, 1972).
[23]H. H. Schaefer, *Topological Vector Spaces* (Springer-Verlag, New York, 1970).

A posterior Schrödinger equation for continuous nondemolition measurement

V. P. Belavkin[a)]
Institute of Physics, Copernicus University, Toruń, Poland

(Received 28 February 1989; accepted for publication 21 March 1990)

A continuous model for a nondemolition observation of an atom is given. An equation for the corresponding instrument is found and a stochastic dissipative Schrödinger equation for the unnormalized posterior wave function of the atom is derived. It is shown that the continuously observed isolated atom relaxes to the ground state without mixing.

I. INTRODUCTION

The time evolution of a quantum system, say atoms, under a continuous observation of photon emission can be obtained in the framework of quantum stochastic theory of nondemolition measurements recently developed in Refs. 1–5. This theory is based on the quantum stochastic counting method worked out by Hudson and Parthasarathy for quantum processes in Ref. 6 and the notion of output quantum fields introduced by Gardiner and Collet.[7] Under the assumption of completeness of the nondemolition observation of the system, we give a direct derivation of a new stochastic linear dissipative wave equation, obtained in Ref. 8, by using a quantum filtering nonlinear equation, called a posterior Schrödinger equation in the case of a continuous measurement, if the observed information is taken into account. As it is shown in Ref. 9, the equation of such a type describes the continuous nonmixing collapse of the wave packet whose propagation depends on the measurement data up to the present instant of time $t>0$. It allows us to explain the Zeno paradox for a free quantum particle by finding a watchdog effect for the posterior wave packet.[9] In this paper we apply the derived equation for the description of the posterior relaxation of an atom continuously observed by measuring the emitted photon field. In contrast to the prior dynamics, the posterior dynamics for the complete observation of the output Boson field gives the relaxation without the mixing of any quantum state of the finite-level atom to its ground state. The impossibility to obtain this intuitively obvious result from the usual Schrödinger equation for an isolated atom can be considered as a quantum paradox of the Zeno kind if one ignores the perturbation of the quantum dynamics under an observation, which is described in this paper.

II. A MODEL OF CONTINUOUS OBSERVATION

Let us consider a quantum system, i.e., "atom + one-dimensional Bose field," the unitary evolution of which satisfies in the singular coupling limit the quantum stochastic Schrödinger equation in the Ito sense:[6,7]

$$dU + KU\,dt = (L \otimes dB^{+} - L^{+} \otimes dB)U, \quad U(0) = I. \tag{1}$$

Here $K = L^{+}L/2 + iH/\hbar$, H is a Hamiltonian of the atom, and L is some operator of the atom coupled to the Bose noise $B(t)$, defining together with their Hermitian conjugates $L^{+}, B^{+}(t)$ the "interaction Hamiltonian"

$$H_{\text{int}} = i\hbar(L \otimes dB^{+} - L^{+} \otimes dB)/dt.$$

The noise is described by creation and annihilation processes,

$$B^{+}(t) = \int_0^t b^{+}(r)dr, \quad B(t) = \int_0^t b(r)dr, \tag{2}$$

where $[b(t'), b^{+}(t)] = \delta(t-t') = \langle b(t')b^{+}(t)\rangle$. The last means that the initial state is taken as the product of a state for the atom and the vacuum state of the quantum noise.

In the case of photon emission, one should take L as an operator proportional to the annihilation operator A in the energy representation $H = \Sigma_{i=1}^{\infty} \epsilon_m |m\rangle\langle m|$ of the atom:

$$L = \sqrt{\lambda}\,A, \quad A = \sum_{m=1}^{\infty} \sqrt{m}\,|m-1\rangle\langle m|,$$

$$A^{+} = \sum_{m=1}^{\infty} \sqrt{m}\,|m\rangle\langle m-1|, \tag{3}$$

where $|m\rangle$ is an eigenvector of the operator H, having the discrete spectrum.

The output Bose field[7] is described by the annihilation process that for all $s \geqslant t$ remains unchanged in the Heisenberg picture:[10]

$$\hat{B}(t) = U^{+}(t)B(t)U(t) = U^{+}(s)B(t)U(s), \quad s \geqslant t.$$

It means that the process $\hat{B}$, as well as $\hat{B}^{+}$, satisfy the nondemolition principle[4,5]

$$[\hat{B}(r), \hat{X}(t)] = U^{+}(t)[B(r), X]U(t) = 0, \quad r \leqslant t, \tag{4}$$

with respect to any operator X of the atom in the Heisenberg picture $\hat{X}(t) = U^{+}(t)XU(t)$. As it is proven in Ref. 5, the condition (4) gives us the possibility to define the posterior mean values of $\hat{X}(t)$ under the condition of observation of any nonanticipating function of $\hat{B}$ and $\hat{B}^{+}$ up to the moment t.

Let us consider the continuous measurement of the output field coordinate process

$$\hat{Q}(t) = \hat{B}(t) + \hat{B}^{+}(t) = U^{+}(s)Q(t)U(s), \quad s \geqslant t,$$

where $Q = B + B^{+}$ is the input Wiener process. Due to the commutativity of the Hermitian process $Q(t)$ with $Q(t')$ for

[a)] On leave of absence from M.I.E.M., 109028 Moscow, USSR.

 0022-2488/90/122930-05$03.00

any t, t', the output nondemolition process $\hat{Q}$ is self-nondemolition i.e., nondemolition in the sense of[11]

$$[\hat{Q}(t'),\hat{Q}(t)]=0, \quad \forall t,t'. \tag{5}$$

Also, the exponential output process

$$\hat{Y}(l,t)=\exp\left\{\int_0^t l(r)d\hat{Q}(r)\right\},$$

is nondemolition and self-nondemolition for any integrable c-valued function $l(t)$.

Let us find the equation for the generating map

$$\Gamma(l,t):X\mapsto\Gamma(l,t)[X],$$

$$\Gamma(l,t)[X]=\int \exp\left\{\int_0^t l(r)dq(r)\right\}J^t(dq)[X], \tag{6}$$

of the corresponding instrument J^t[12] on the algebra of operators X of the atom. By the definition of the instrument it defines, in the Schrödinger picture, the time evolution $\rho\mapsto\rho^t(dq)$ of an initial state $\rho:X\mapsto\rho[X]$ of the atom to the state $\rho^t(dq)=\rho\circ J^t(dq)$, normalized on the probability $\mu^t(dq')=\rho[J^t(dq)[I]]$ of the observed event dq up to the instant $t>0$. The generating map can be found from the condition

$$\langle\psi|\Gamma(l,t)[X]\psi\rangle=\langle\hat{Y}(l,t)\hat{X}(t)\rangle, \tag{7}$$

where the mean value $\langle\cdot\rangle$ is taken with respect to the initial pure state being the product of the wave function ψ of the atom and the vacuum state vector of the noise. To this aim one should find the quantum stochastic equation for $\hat{G}=\hat{Y}\hat{X}$, using the Ito formula

$$d(\hat{Y}\hat{X})=d\hat{Y}\hat{X}+\hat{Y}d\hat{X}+d\hat{Y}d\hat{X}, \tag{8}$$

and Heisenberg quantum stochastic Ito equations[6]

$$d\hat{X}+(\hat{K}^+\hat{X}+\hat{X}\hat{K}-\hat{L}^+\hat{X}\hat{L})dt$$
$$=[\hat{X},\hat{L}]dB^+ +[\hat{L}^+,\hat{X}]dB, \tag{9a}$$

$$d\hat{Y}=l\hat{Y}d\hat{Q}+\tfrac{1}{2}l^2\hat{Y}dt, \quad d\hat{Q}=(\hat{L}+\hat{L}^+)dt+dQ, \tag{9b}$$

$$\hat{K}(t)=U^+(t)KU(t), \quad \hat{L}(t)=U^+(t)LU(t).$$

We shall do it directly by applying the Ito formula for the product $\hat{G}=U^+GU$, where

$$G(t)=Y(t)X, \quad Y(l,t)=\exp\left\{\int_0^t l(r)dQ(r)\right\},$$

obtaining (9) for $l=0$, $G=X$, and $X=I$, $G=Y$, respectively. Taking into account that

$$dG=lG\,dQ+\tfrac{1}{2}l^2G\,dt, \quad \text{for } G(t)=X\exp\left\{\int_0^t l\,dQ\right\},$$

we have from (1) and the multiplication rules

$$dB\,dB^+=dt, \quad (dB)^2=dB^+\,dB=(dB^+)^2=0,$$

$$d\hat{G}=dU^+GU+U^+dG\,U+U^+G\,dU+dU^+dG\,U$$
$$+dU^+G\,dU+U^+dG\,dU$$
$$=U^+(L^+G+G(l-L^+))U\,dB$$
$$+U^+((l-L)G+GL)U\,dB^+$$
$$+U^+(\tfrac{1}{2}l^2G-K^+G-GK+L^+Gl$$
$$+L^+GL+lGL)U\,dt.$$

It helps to write the equation for mean value (7) as

$$d\langle\hat{G}(t)\rangle=\langle\hat{\eta}(t)|\tfrac{1}{2}l^2G-K^+G-GK$$
$$+l(L^+G+GL)+L^+GL|\hat{\eta}(t)\rangle dt,$$

where $\hat{\eta}(t)=U(t)\eta$, $\eta=\psi\otimes\delta_\phi$, and δ_ϕ is the vacuum vector of the noise:

$$dB(t)U(t)\eta=U(t)dB(t)\eta=0,$$

for $dB(t)=B(t+dt)-B(t)$.

Hence, the generating map Γ defined in (7), satisfies the forward differential equation

$$\frac{d}{dt}\Gamma[X]=\Gamma\left[\frac{1}{2}l^2X-K^+X-XK\right.$$
$$\left.+l(L^+X+XL)+L^+XL\right], \tag{10}$$

with the initial condition $\Gamma(l,0)[X]=X$.

III. A POSTERIOR QUANTUM DYNAMICS

Now we shall obtain the solution of Eq. (10) in the form

$$\Gamma(l,t)[X]=\int_{\Omega^t}Y(l,q^t)V^*(q^t)XV(q^t)d\nu(q^t), \tag{11}$$

where ν is the standard Wiener probability measure on the space Ω of continuous trajectories $q=[q(t)|t>0]$ of the observed process $\hat{Q}$, restricted to the space $\Omega^t=\{q^t|q\in\Omega\}$ of the trajectories stopped at t: $q^t=[q(r)|r\leqslant t]$, and $Y(l,q^t)=\hat{Y}(l,t)(q)$. Comparing (6) and (11) and taking into account that $\hat{Y}(l,t)(q)=\exp\int_0^t l(r)dq(r)$ one can represent the instrument $J^t(dq):X\mapsto J(dq^t)[X]$ as an absolutely continuous one with respect to $d\nu(q)$:

$$J(dq^t)[X]=V^+(q^t)XV(q^t)d\nu(q^t). \tag{12}$$

In order to prove it, let us identify the standard Wiener process with the input process Q, having the standard measure on its spectrum Ω with respect to the vacuum state. We shall show that the stochastic propagator $\hat{V}(t)(q)=V(q^t)$, defining for any trajectory $q=[q(t)]$ the posterior evolution $V(q^t):\psi\mapsto V(q^t)\psi=\hat{\chi}(t)(q)$ for the unnormalized stochastic wave function $\hat{\chi}(t)$ of the atom, satisfies the stochastic Schrödinger equation

$$dV+KV\,dt=LV\,dQ, \quad V(0)=I. \tag{13}$$

Indeed, if V satisfies Eq. (13) in the Ito sense, then

$$d(V^+XV)=dV^+XV+V^+X\,dV+dV^+X\,dV$$
$$=V^+(L^+X+XL)V\,dQ$$
$$-V^+(K^+X+XK-L^+XL)V\,dt.$$

It gives the recursive filtering equation

$$d\Phi[X]+\Phi[K^+X+XK-L^+XL]dt$$
$$=\Phi[L^+X+XL]dQ, \tag{14}$$

$\Phi(0)[X]=X$ for the stochastic map,

$$\Phi(t)[X]=V^+(t)XV(t), \quad \Phi(t)(q)=\Phi(q^t), \tag{15}$$

defining a selective instrument $\Phi(q^t)[X]=V^+(q^t)XV(q^t)$ for any trajectory $q=[q(t)]$. Taking into account that

$$d(Y(l)\Phi[X])$$
$$= dY(l)\Phi[X] + Y(l)d\Phi[X] + dY(l)d\Phi[X]$$
$$= Y(l)\Phi[lX + L^{+}X + XL]dQ$$
$$+ Y(l)\Phi[\tfrac{1}{2}l^2X - K^{+}X - XK$$
$$+ l(L^{+}X + XL) + L^{+}XL]dt,$$

and averaging it with respect to the standard Wiener measure one obtains (10) for the mean value (12) of the product $Y(l,q^t)\Phi(q^t)[X]$. So, the wave function $\hat{\chi}(t) = V(t)\psi$, satisfies the stochastic dissipative equation

$$d\hat{\chi} + (L^{+}L/2 + iH/\hbar)\hat{\chi}\,dt = L\hat{\chi}\,d\hat{Q}, \quad \hat{\chi}(0) = \psi, \tag{16}$$

with respect to the output process $\hat{Q}$, coinciding with Q in the Schrödinger picture. It is normalized on the probability density

$$p(q^t) = \langle V(q^t)\psi | V(q^t)\psi\rangle \equiv \hat{p}(t)(q), \tag{17}$$

of the observed process $\hat{Q}$ with respect to the standard Wiener measure of the input process Q. It follows from the integral representation that

$$\langle \hat{Y}(l,t)\hat{X}(t)\rangle$$
$$= \int_{\Omega^t} Y(l,q^t)\langle V(q^t)\psi | XV(q^t)\psi\rangle d\nu(q^t)$$
$$= \int_{\Omega^t} Y(l,q^t)p(q^t)\langle X\rangle(q^t)d\nu(q^t) \tag{18}$$

of (7), giving for $X = I$ the mean value of (6) for the exponential output process (6) as the generating function of the output probability measure

$$d\mu(q^t) = p(q^t)d\nu(q^t). \tag{19}$$

The formula (18) defines the posterior mean value $\langle X\rangle(q^t)$ as

$$\langle X\rangle(q^t) = \langle \varphi(q^t)|X\varphi(q^t)\rangle \equiv \langle X\rangle^t(q),$$

in terms of the normalized posterior wave function $\hat{\varphi}(t)(q) = \varphi(q^t)$, $\varphi(q^t) = \chi(q^t)/p(q^t)^{1/2}$ satisfying the nonlinear stochastic wave equation

$$d\hat{\varphi} + (\tilde{L}^{+}\tilde{L}/2 + i\tilde{H}/\hbar)\hat{\varphi}\,dt = \tilde{L}\hat{\varphi}\,d\tilde{Q}, \tag{20}$$

where $\tilde{L}(t) = L - \mathrm{Re}\langle L\rangle^t$, $\tilde{H}(t) = H - \hbar\,\mathrm{Re}\langle L\rangle^t \mathrm{Im}\,L$, $d\tilde{Q}(t) = d\hat{Q}(t) - 2\,\mathrm{Re}\langle L\rangle^t dt$, obtained in Ref. 8 from the filtering equation derived in Ref. 4. It can be found directly from (16) by using the Ito formula for $\hat{\chi} = \hat{c}\hat{\varphi}$: $d\hat{\chi} = d\hat{c}\,\hat{\varphi} + \hat{c}\,d\hat{\varphi} + d\hat{c}\,d\hat{\varphi}$, where $d\hat{c}\,d\hat{\varphi} = \tilde{L}\,\mathrm{Re}\langle L\rangle^t\hat{\chi}\,dt$, for $\hat{c}(t) = (\hat{\chi}(t)^{+}\hat{\chi}(t))^{1/2}$, satisfying the stochastic equation

$$d\hat{c} + (\mathrm{Re}\langle L\rangle^t)^2\hat{c}\,dt/2 = \mathrm{Re}\langle L\rangle^t\hat{c}\,d\hat{Q}, \quad \hat{c}(0) = 1.$$

This gives a Girsanov transformation $\hat{Q}\to\tilde{Q}$ to another Wiener process $\tilde{Q}$ with respect to (19) for $p(q^t) = |\hat{c}(t)|^2(q)$.

IV. AN OBSERVATION OF PHOTON EMISSION

Let us consider the stochastic model of a continuous observation of photon emission described by the coupling operator $L = \sqrt{\lambda}\,A, \lambda > 0$, defined in (3) in the energy representation of the atom. The linear posterior wave equation (16)

$$d\hat{\chi} + (\lambda A^{+}A/2 + iH/\hbar)\hat{\chi}\,dt = \sqrt{\lambda}\,A\hat{\chi}\,d\hat{Q},$$

gives in the energy representation the following chain of stochastic equations:

$$\hat{d}x_m + \varkappa_m\hat{x}_m\,dt = \sqrt{(m+1)\lambda}\,\hat{x}_{m+1}\,d\hat{Q}, \quad \hat{x}_m(0) = c_m, \tag{21}$$

for the coefficients x_m of the expansion

$$\hat{\chi}(t) = \sum_{m=0}^{\infty} |m\rangle\hat{x}_m(t), \quad \psi = \sum_{m=0}^{\infty} |m\rangle c_m.$$

The complex parameters $\varkappa_m = m\lambda/2 + i\epsilon_m/\hbar$ appearing in (21) are the eigenvalues of the normal operator $K = \lambda A^{+}A/2 + iH/\hbar$, corresponding to the eigenvectors $|m\rangle$, for which $A|m\rangle = \sqrt{m}|m-1\rangle$ holds. The chain (21) is finite if the initial state $\hat{\chi}(0) = \psi$ has zero coefficients $c_m = 0$, $m \geqslant M$ for some $M < \infty$, as it always happens in the case of the M-level atom.

Let us show that the solution of the system (21) can be written in terms of the multiplicative stochastic integral on the set Ω^t of all finite chains $\tau = (t_1,...,t_l), l = 0,1,..., 0 < t_1 < \cdots < t_l \leqslant t$, as

$$\hat{x}_m(t) = e^{-\varkappa_m t}\int_{\Omega^t}\left(\lambda^{|\tau|}\frac{(m+|\tau|)!}{m!}\right)^{1/2} c_{m+|\tau|}$$
$$\times e^{-\alpha_m(\tau)}\,d\hat{Q}(\tau), \tag{22}$$

for $m = 0,1,...$. In (22) $|\tau|$ denotes the length $l = \Sigma_{t\in\tau}\chi(t)$ of the chain τ, $\alpha_m(\tau) = \Sigma_{t\in\tau}(\lambda/2 + i\omega_{m+n(t)})t$, $\omega_n = \varepsilon_n - \varepsilon_{n-1}$, $n(t) = \Sigma_{s\in\tau}\chi(s-t)$ — the number of t in the chain $\tau = (t_1,...,t_l)$ counted from its end, $dQ(\tau) = \Pi_{t\in\tau}\,dQ(t)$, $\Pi_{t\in\tau}\,dQ(t) = dQ(t_1)\cdots dQ(t_l)$ for $\tau = (t_1,...,t_l)$, and for any integrable function of f of $\tau \in \Omega^t$ $\int_{\Omega^t} f(\tau)dQ(\tau)$ denotes the sum of the repeated stochastic integrals

$$\int_{\Omega^t} f(\tau)d\hat{Q}(\tau)$$
$$= \sum_{l=0}^{\infty}\int_{t>t_l>}\cdots\int_{>t_1>0} f(t_1,...,t_l)\prod_{i=1}^{l} d\hat{Q}(t_i). \tag{23}$$

Indeed, representing the integral (23) for

$$f_m(\tau) = (\lambda^{|\tau|}(m+|\tau|)!/m!)^{1/2}c_{m+|\tau|}\,e^{-\alpha_m(\tau)},$$

as

$$\int_{\Omega^t} f(\tau)d\hat{Q}(\tau) = f(0) + \int_0^t d\hat{Q}(r)\int_{\Omega^r} f(\tau,r)d\hat{Q}(\tau),$$

where $f(0) = c_m$ for the chain $\tau = 0$ of the length $l = |0| = 0$, and

$$f_m(\tau,r) = \sqrt{\lambda(m+1)}f_{m+1}(\tau)\exp\{-(\lambda/2 + i\omega_{m+1})r\}.$$

Taking into account that $\varkappa_{m+1} - \varkappa_m = \lambda/2 + i\omega_{m+1}$, one can rewrite (22) in the form of the recursive stochastic equation

$$\hat{x}_m(t)$$
$$= e^{-\varkappa_m t}c_m - \int_0^t e^{\varkappa_m(r-t)}\sqrt{\lambda(m+1)}\hat{x}_{m+1}(r)d\hat{Q}(r). \tag{24}$$

This gives the system of stochastic integral equations

$$\hat{z}_m(t) = c_m + \int_0^t \sqrt{(m+1)\lambda}\,e^{(\varkappa_m - \varkappa_{m+1})r}\hat{z}_{m+1}(r)d\hat{Q}(r),$$

for $\hat{z}_m(t) = e^{-\varkappa_m t}\hat{x}_m(t)$, having the differential form

$$d\hat{z}_m(t) = \sqrt{(m+1)\lambda}\, e^{(\varkappa_m - \varkappa_{m+1})t}\hat{z}_{m+1}(t)d\hat{Q}(t), \qquad (25)$$

equivalent to the system of stochastic differential equations (21) for $\hat{x}_m(t) = e^{-\varkappa_m t}\hat{z}_m(t)$:

$$d\hat{x}_m = e^{-\varkappa_m t}\, d\hat{z}_m(t) - \varkappa_m \hat{x}_m(t)dt$$
$$= \sqrt{(m+1)\lambda}\,\hat{x}_{m+1}\, d\hat{Q} - \varkappa_m \hat{x}_m\, dt.$$

Let us now find the conditional probability $p_m(t\,|q) = p_m(q^t)$ for the atom to be on the mth level after the nondemolition observation up to t defined by the probability density $|\hat{x}_m(t)|^2(q) = |x_m(q^t)|^2$ with respect to the standard Wiener probability measure ν on Ω^t:

$$p_m(q^t) = |x_m(q^t)|^2\Big(\sum_{m=0}^{\infty}|x_m(q^t)|^2\Big)^{-1}. \qquad (26)$$

In order to do this we need the multiplication formula[13] for Wiener integrals (23):

$$\left|\int_{\Omega^t} f(\tau)d\hat{Q}(\tau)\right|^2$$
$$= \int_{\Omega^t}\Big(\sum_{\rho\cup\sigma=\nu}\int_{\Omega^t} f(\tau\cup\rho)^* f(\tau\cup\sigma)d\tau\Big)d\hat{Q}(\nu), \qquad (27)$$

where the sum is taken over partitions $\rho\cup\sigma = \rho\cup\sigma$, $\rho\cap\sigma = \phi$ of the chains $\nu = (v_1,...,v_n)\in\Omega^t$ into the chains $\rho = (r_1,...,r_j)$, $\sigma = (s_1,...,s_k)$ with the total length $j+k = n$. Using this formula for the stochastic integral (22) one obtains

$$|\hat{x}_m(t)|^2 = e^{-\lambda m t}\int_{\Omega^t}\lambda^{|\tau|}e^{-|\lambda|\bar{\tau}}\hat{p}^t_m(\tau)d\tau, \qquad (28)$$

where

$$\bar{t} = \sum_{t\in\tau} t/|\tau|, \quad d\tau = \prod_{t\in\tau} dt \quad \Big[\bar{\tau} = \sum_{n=1}^{l}\frac{t_n}{l},$$
$$d\tau = \prod_{n=1}^{l} dt_n, \text{ if } \tau = (t_1,...,t_l)\Big]$$
$$\hat{p}^t_m(\tau) = \frac{1}{m!}\int_{\Omega^t}\lambda^{|\nu|/2}e^{-|\nu|\lambda\bar{v}/2}$$
$$\times\sum_{\rho\cup\sigma=\nu} c^{(m)}(\sigma\cup\tau)^* c^{(m)}(\rho\cup\tau)dQ(\nu), \qquad (29)$$
$$c^{(m)}(\tau) = \sqrt{(m+|\tau|)!}\exp\Big\{i\sum_{t\in\tau}\omega_{m+n(t)}t\Big\}c_{m+|\tau|},$$
$$\bar{v} = \sum_{v\in\nu}\frac{v}{|\nu|} \quad \Big[\bar{v} = \sum_{l=1}^{n}\frac{v_l}{n}, \quad \text{if } \nu = (v_1,...,v_n)\Big]. \qquad (30)$$

V. A POSTERIOR QUANTUM RELAXATION

As it follows from (28), the probability (26) for the electron in the continuously observed atom to be on the mth level, $m>0$, decays exponentially to zero for $t\to\infty$ if

$$\hat{s}_m = \int_{\Omega}\lambda^{|\tau|}e^{-|\tau|\lambda\bar{\tau}}\hat{p}_m(\tau)d\tau < \infty, \qquad (31)$$

where $\hat{p}_m(\tau) = \lim \hat{p}^t_m(\tau)$ for $t\to\infty$ and the integral is taken over the space Ω of all the chains $\tau = (t_1,...,t_l)$, $0<t_1<\cdots<t_l<\infty$ of finite length $l = 0,1,...$. Equation (31) obviously holds for the case of the finite-level atom for which the integral (31) is the finite sum,

$$\bar{s}_m = \sum_{l=0}^{M-1}\lambda^l\int_{\infty>t_l>}\cdots\int_{>t_1>0}\exp-\Big(\lambda\sum_{n=1}^{l}t_n\Big)$$
$$\times\hat{p}_m(t_1,...,t_l)\prod_{i=1}^{l}dt_i, \qquad (32)$$

of the repeated integrals of the stochastic functions

$$\hat{p}_m(\tau) = \frac{1}{m!}\sum_{n=0}^{M-m-|\tau|}\int_{\infty>v_n>}\cdots\int_{>v_1>0}\lambda^{n/2}$$
$$\times\exp\Big(-\lambda\sum_{l=1}^{n}\frac{v_l}{2}\Big)\sum_{\rho\cup\sigma=(v_1,\cdots,v_n)} c^{(m)}(\sigma\cup\tau)^*$$
$$\times c^{(m)}(\rho\cup\tau)\prod_{l=1}^{n}dQ(v_l),$$

having the zero values $\hat{p}_m(\tau) = 0$ for $l = |\tau|\geqslant M-m$ due to $c_{m+|\tau|} = 0$, and being the finite sums of repeated stochastic integrals for $m+|\tau| < M$. It means that the finite-level atom relaxes under a continuous observation of photon emission to the ground state $|0\rangle$ without mixing.

Let us prove this posterior relaxation also for the infinite-level atom provided that the initial state $\Sigma\, c_m|m\rangle$ satisfies the conditions

$$|c_m|\leqslant(\hat{K}/m!)^{1/2}|z|^m, \quad m = 0,1,2,..., \qquad (33)$$

where $\hat{K}>0$. The conditions (33), fulfilled obviously in the finite-level case ($c_m = 0, m\geqslant M$), is also satisfied for a linear combination of the Poissonian (coherent) amplitudes

$$c_m = z^m\exp\{-\tfrac{1}{2}|z|^2\}/(m!)^{1/2}, \quad m = 0,1,2,... .$$

In the case (30) we obtain the following estimation for (29):

$$\hat{p}^t_m(\tau)\leqslant\frac{\hat{K}}{m!}|z|^2\int_{\Omega^t}\lambda^{|\nu|/2}e^{-|\nu|\lambda\bar{v}/2}(2|z|)^{|\nu|}d\hat{Q}(\nu), \qquad (34)$$

where we have taken into account that

$$|c^{(m)}(\tau)|\leqslant\hat{K}^{1/2}|z|^{m+|\tau|}, \quad \sum_{\sigma\cup\rho=\nu}|z|^{|\sigma|}|z|^{|\rho|} = (2|z|)^{|\nu|}.$$

The integrand on the right-hand side of (34) is a product function

$$\lambda^{|\nu|/2}e^{-|\nu|\lambda\bar{v}/2}(2|z|)^{|\nu|} = \prod_{v\in\nu}\sqrt{\lambda}\,2|z|e^{-\lambda v/2}.$$

Therefore one can use the stochastic exponential formula[13]

$$\int_{\Omega^t}\prod_{v\in\nu}f(v)d\hat{Q}(v) = \exp\Big\{\int_0^t f(v)d\hat{Q}(v) - \frac{1}{2}\int_0^t f(v)^2dv\Big\}, \qquad (35)$$

for the Wiener integral $\int_{\Omega^t} f(\nu)d\hat{Q}(\nu)$ with $f(\nu) = \Pi_{v\in\nu}f(v)$. Due to that, the inequality (34) takes the form

$$\hat{p}^t_m(\tau)\leqslant\frac{\hat{K}}{m!}|z|^{2|\tau|}\exp\Big\{\int_0^t\sqrt{\lambda}\,2|z|e^{-\lambda v/2}d\hat{Q}(v)$$
$$-2|z|^2(1-e^{-\lambda t})\Big\}. \qquad (36)$$

Now we can estimate the integral in (28):

$$\hat{s}_m^t = \int_{\Omega^t} \lambda^{|\tau|} e^{-|\tau|\lambda\bar{t}}\, \hat{p}_m^t(\tau)\, d\tau$$

$$\leqslant \left\{ \frac{\hat{K}}{m!} \exp\left\{ \int_0^t \sqrt{\lambda}\, 2|z| e^{-\lambda v/2}\, d\hat{Q}(v) - 2|z|^2(1-e^{-\lambda t}) \right\} \right\} \int_{\Omega^t} (|z|^2\lambda)^{|\tau|} e^{-|\tau|\lambda\bar{t}}\, d\tau$$

$$= \frac{\hat{K}}{m!} \exp\left\{ \int_0^t \sqrt{\lambda}\, 2|z| e^{-\lambda v/2}\, d\hat{Q}(v) - |z|^2(1-e^{-\lambda t}) \right\},$$

where we have taken into account the exponential formula

$$\int_{\Omega^t} \prod_{\tau \in \tau} f(t)\, dt = \sum_{l=0}^{\infty} \int_{t>t_l} \cdots \int_{>t_1>0} \prod_{n=1}^{l} f(t_n)\, dt_n = \exp\left\{ \int_0^t f(r)\, dr \right\}, \tag{37}$$

for $f(t) = |z|^2\lambda \exp\{-\lambda t\}$. So we obtain the estimation

$$|\hat{x}_m(t)|^2 \leqslant e^{-\lambda m t} \frac{\hat{K}}{m!} \exp\left\{ 2\sqrt{\lambda}\,|z| \int_0^t e^{-\lambda v/2}\, d\hat{Q}(v) - |z|^2(1-e^{\lambda t}) \right\} \tag{38}$$

for the probability of the observed atom to be on the mth level, normalized on the probability density $\hat{p}(t) = \Sigma_{m=0}^{\infty} |\hat{x}_m(t)|^2$ for the observation of $\hat{Q}$ with respect to the standard Wiener measure. For $\hat{p}(t)$ we obtain

$$\hat{p}(t) \leqslant \hat{K} \exp\left\{ 2\sqrt{\lambda}\,|z| \int_0^t e^{-\lambda v/2}\, d\hat{Q}(v) + e^{-\lambda t}(|z|^2+1) - |z|^2 \right\}. \tag{39}$$

Hence, $|\hat{x}_m(t)|^2 = e^{-\lambda m t}\hat{s}_m^t$ for $m \neq 0$, decays exponentially to zero as $t \to \infty$ with the rate

$$\hat{s}_m = \lim \hat{s}_m(t) \leqslant \frac{\hat{K}}{m!} \exp\left\{ 2\sqrt{\lambda}\,|z| \int_0^{\infty} e^{-\lambda v/2}\, d\hat{Q}(v) - |z|^2 \right\}. \tag{40}$$

Equation (16) which appeared in Ref. 8 first, was obtained in the nonlinear form (20) in Refs. 3 and 4 and in the general density matrix form in Ref. 14. Some particular cases of these posterior equations has been postulated recently in connection with the stochastic models of the dynamical theory of wave function reduction.[15–17]

[1] V. P. Belavkin, Radiotech. Elektron. **25**, 1445 (1980).
[2] A. Barchielli and G. Lupieri, J. Math. Phys. **26**, 2222 (1985).
[3] V. P. Belavkin, in *Modelling and Control of Systems*, edited by A. Blaquiere, Lecture Notes in Control and Inform. Sciences, Vol. 121 (Springer, Berlin, 1988), p. 245.
[4] V. P. Belavkin, in *Stochastic Methods in Mathematics and Physics*, Proc. XXIV Karpacz winter school, edited by R. Gielerak and W. Karwowski (World Scientific, Singapore, 1989), p. 310.
[5] V. P. Belavkin, "Quantum stochastic calculus and quantum nonlinear filtering," Preprint Centro Matematico V. Volterra, Univ. di Roma II n.6 (1989).
[6] R. L. Hudson and K. R. Parthasarathy, Commun. Math. Phys. **93**, 301 (1984).
[7] C. W. Gardiner and M. J. Collett, Phys. Rev. A **31**, 3761 (1985).
[8] V. P. Belavkin, "A new wave equation for a continuous nondemolition measurement," Phys. Lett. A **140**, 355 (1989).
[9] V. P. Belavkin and P. Staszewski, "A quantum particle undergoing continuous observation," Phys. Lett. A **140**, 359 (1989).
[10] V. P. Belavkin, in *Quantum Probability and Applications V*, edited by L. Accardi and W. von Waldenfels (Springer, New York, 1990).
[11] C. M. Caves, in: *Quantum Optics, Experimental Gravitation, and Measurement Theory*, edited by P. Meystre and M. D. Scully (Plenum, New York, 1982).
[12] E. B. Davies, *Quantum Theory of Open Systems* (Academic, New York, 1976).
[13] M. Lindsay and H. Maassen, in *Quantum Probability and Applications III*, edited by L. Accardi and W. von Waldenfels (Springer, New York, 1988), p. 192.
[14] V. P. Belavkin "A Boson field stochastic calculus and quantum filtering in Fock space," Preprint 723, Dipartamento di Fisica, Universita di Roma, La Sapienza (1989).
[15] L. Diosi, Phys. Lett. A **129**, 419; **132**, 233 (1988).
[16] N. Gisin, Helv. Phys. Acta **62**, 363 (1989).
[17] G. C. Ghivardi, A. Rimini, and P. Pearle, Trieste preprint IC/89/44, to appear in Phys. Rev. A.

Corrections to the coherent state path integral: Comments upon a speculation of L. S. Schulman

T. L. Marchioro II[a)]
Department of Physics, University of Washington, Seattle, Washington 98195

(Received 16 April 1990; accepted for publication 8 August 1990)

Intuitively, the Feynman path integral corresponds to a weighted sum over classical paths, an interpretation that fails for the phase space path integral. To address the question of whether there exists a path integral expression conforming to a sum over paths in phase space, an examination of the discrete coherent state path integral (CSPI) is undertaken. Via an alternative formulation of the CSPI, it is shown that the coherent state action for a broad class Hamiltonians can be transformed from the variables (q,p) to $(q,\dot{q})$. For these Hamiltonians, such a transformation along with the inclusion of all terms $O(\epsilon(z_i - z_{i-1}))$ yields an expression which, for finite ϵ, can be interpreted as a sum over classical paths with Gaussian weight. The numerical evaluation of this expression through importance sampling (Monte Carlo) is demonstrated.

I. INTRODUCTION

Feynman's path integral formulation of quantum mechanics[1] has enjoyed enormous appeal for nearly 40 years. The reasons for this are myriad, but perhaps greatest among them is that the path integral, while an elegantly powerful formal tool, also admits to a relatively simple intuitive interpretation; that is, a weighted sum over all possible classical paths. This interpretation proves particularly attractive in providing insight into the semiclassical approximation of quantum mechanics, which, in the path integral formalism, yields the classical path as a contour of stationary phase. This heuristic "sum over histories" interpretation derives from the fact that, for virtually all cases in which the Feynman path integral (time-dependent Green's function) can be analytically evaluated, the resulting expression corresponds to a sum over purely classical paths.[2]

Thus considerable attention has been devoted to providing a rigorous definition of the propagator as a true integral in the space of paths (that is, in the manner of a Wiener integral[3]), both because this would verify the intuitive interpretation of the path integral and also because it would elevate Feynman's heuristic expression to the status of a true functional integral. This process essentially requires interchanging the orders of integration and the $N\to\infty$ $(\epsilon\to 0)$ limit in the discrete time slice approximation of the path integral. For the specific case of the Feynman path integral, this quest has proven most problematic; it has not been possible to specify a well-defined measure for a Wiener integral with complex diffusion constant.[2,4] Whether this difficulty can be circumvented remains, to my knowledge, an open question.

Nonetheless, the larger goal of a well-defined path integral expression for the quantum propagator continues to invite attention. In this regard, two other path integral formulations of quantum dynamics have been investigated, the phase space path integral and the coherent state path integral (henceforth, the Feynman, phase space, and coherent state path integrals will be denoted by FPI, PSPI, and CSPI, respectively). Both of the these formulations are based upon summing paths in phase space and thus prove more general than the FPI which integrates only over configuration space. For the PSPI attempts at a rigorous formulation fail even more abysmally than with the FPI;[2] for, while the FPI sums over paths that are continuous but nowhere differentiable, the phase space "paths" prove neither continuous nor differentiable. In fact, the most suggestive aspect of the PSPI studies is that the process of transforming the PSPI into the FPI [essentially a functional Legendre transform from the variables (q,p) to $(q,\dot{q})$] significantly smooths the paths of integration.

For the CSPI, the results have proven more promising. The localized nature of the coherent states has long suggested that only a small portion of the phase space contributes to the dynamics[5] (i.e., the space of paths has compact support). Moreover, because the overcompleteness of the coherent state basis implies the CSPI is not unique (over and above questions of operator ordering—see below), there may exist alternate formulations of the CSPI for which the sum over paths interpretation is more transparent. Indeed, Klauder and Daubechies[6] recently presented a form of the CSPI interpretable as an expression involving phase space paths of Wiener measure in the limit as the diffusion constant diverges, thus implying a regularity to the CSPI paths not present in either the FPI or the PSPI.

From the above arguments one could logically conclude that if a well-defined path integral expression for the quantum propagator exists, it will be found in the framework of the CSPI. Taking this conclusion as a starting point, this paper re-examines the CSPI. An alternative approach is taken which involves (1) for Hamiltonians of the form $p^2/2m + V(x)$ a Legendre transform of the coherent state action to the variables $(q,\dot{q})$, and (2) retention of all terms of the form $\epsilon(z_i - z_{i-1})$ in the coherent state action.

The Legendre transform allows the momentum integrals to be performed analytically leaving an expression in configuration space only. In analogy with the PSPI to FPI

[a)] Current address: Department of Chemistry, University of Cincinnati, Cincinnati, OH 45221.

 0022-2488/90/122935-11$03.00

transformation, one would hope the paths of the second-order form action (that is, as functions of q and $\dot{q}$) would be "smoother" than those of the first-order form (functions of q and p). The terms $\epsilon(z_i - z_{i-1})$ are retained because, as was speculated by Schulman,[2] they need not vanish faster than ϵ and can (indeed, as shall be shown, do) contribute to the path integral. The expression resulting from inclusion of these terms and the Legendre transformation has Wiener measure (i.e., Gaussian damping) for finite N, with the damping vanishing for only one integration variable in the path integral ($N \to \infty$) limit. Hence, this expression for the CSPI is both more in the spirit of a true functional integral and, as an added bonus, can be numerically evaluated via importance sampling.

The outline is as follows: Sec. II briefly reviews the formulation of the PSPI, FPI, and CSPI. Section III examines the CSPI for Hamiltonians of the form $H = p^2/2m + V(x)$. A formulation is demonstrated that permits the Legendre transformation of the coherent state action to second-order form. The path integral limit of the corresponding expression is discussed. Section IV retraces the development of Sec. III but with the inclusion of the $\epsilon(z_i - z_{i-1})$ terms. This analysis yields corrections to the conventional CSPI. Section V discusses the significance of these corrections and comments on the properties of the new path integral expression derived.

II. REVIEW OF PATH INTEGRAL FORMULATIONS

The essence of the path integral formulation of quantum dynamics lies in expressing time-dependent quantum transformation amplitudes, through suitable operator factorization and repeated insertions of the identity operator, as the limit of an infinite number of integrations. The FPI, PSPI, and CSPI each accomplish this formal task in a completely general manner. That is, although the FPI and PSPI are commonly presented as means of evaluating coordinate representation transformation functions, and the CSPI as an expression of a coherent state transformation function, any of the three formalisms may be used, for any initial and final state, by introducing at most two extra insertions of unity. The only fundamental difference between the three approaches lies in the integral representation of the identity operator repeatedly inserted, thus leading to an infinite number of different intermediate integration variables with correspondingly different measure. Specifically, the PSPI and CSPI use resolutions of the identity in phase space

$$\hat{I} = \int |p\rangle e^{-ipq/\hbar} \langle q| \frac{dp\,dq}{\sqrt{2\pi\hbar}}$$

and

$$\hat{I} = \int |z\rangle\langle z| \frac{d^2z}{\pi},$$

respectively. The latter representation[7] expresses the completeness of the coherent state basis

$$\hat{a}|z\rangle = |z\rangle z,$$

where

$$z = \frac{1}{\sqrt{2}}\left(\frac{q}{\sigma} + \frac{i\sigma}{\hbar}p\right)$$

and σ is a scale factor with units of length (for the familiar coherent states of a harmonic oscillator $\sigma = \sqrt{\hbar/m\omega}$.)

Feynman's original presentation utilized only the representation of the identity in configuration space

$$\hat{I} = \int |q\rangle\langle q| dq,$$

and the observation, due to Schrödinger and Dirac,[8] that for short time the quantum transformation amplitude

$$\langle q_f|q_i\rangle(t) = \langle q_f|e^{-i\hat{H}t/\hbar}|q_i\rangle \tag{1}$$

was proportional to $e^{iS/\hbar}$, where S is the solution of the corresponding Hamilton–Jacobi equation. From this, Feynman heuristically bootstrapped directly to his original path integral expression. However, as is well known, it is possible to derive the FPI in reasonably rigorous fashion, passing through the PSPI as an intermediate step.

To this end, the time parameter in (1) is divided into equal parts of size $\epsilon = t/(N+1)$ and N insertions of the coordinate representation of the identity are made, yielding

$$\begin{aligned}\langle q_f|q_i\rangle(t) = \int [dq_N \cdots dq_1]\, &\langle q_f|e^{-i\hat{H}\epsilon/\hbar}|q_N\rangle \\ &\times \langle q_N|e^{-i\hat{H}\epsilon/\hbar}|q_{N-1}\rangle \cdots \langle q_2|e^{-i\hat{H}\epsilon/\hbar}|q_1\rangle \\ &\times \langle q_1|e^{-i\hat{H}\epsilon/\hbar}|q_i\rangle. \end{aligned} \tag{2}$$

The time dependence of a typical factor in the integrand (for simplicity the Hamiltonian is assumed to contain no explicit time-dependent piece) may be obtained by inserting a complete set of momentum states and using

$$\begin{aligned}\langle q_j|e^{-i\hat{H}\epsilon/\hbar}|q_{j-1}\rangle &= \int \langle q_j|e^{-i\hat{H}\epsilon/\hbar}|p\rangle\langle p|q_{j-1}\rangle dp \\ &= \int \exp\left[\frac{i}{\hbar}(p(q_j - q_{j-1}) \right. \\ &\quad \left. - \epsilon H(q_j,p))\right]\frac{dp}{2\pi\hbar}. \end{aligned} \tag{3}$$

The replacement of the operators $\hat{p}$ and $\hat{q}$ by their eigenvalues p and q_j assumes a reordering of H through a suitable Trotter product expansion or similar device. These questions of operator ordering will be addressed below; all that matters for the moment is that such an ordering is generally possible but that the c number $H(q,p)$ may differ from simply replacing the operators in $H(\hat{q},\hat{p})$ by their eigenvalues. Carrying out this process on the $N+1$ matrix elements of (2) yields

$$\begin{aligned}\langle q_f|q_i\rangle(t) = \int \frac{dp_{N+1}}{2\pi\hbar}\left[\prod_{k=1}^{N}\frac{dp_k\,dq_k}{2\pi\hbar}\right] &\\ \times \exp\left\{\frac{i}{\hbar}\sum_{k=1}^{N+1} p_k(q_k - q_{k-1})\right. &\\ \left. - \epsilon H(q_k,p_k)\right\}, \end{aligned} \tag{4}$$

where, by definition, $q_{N+1} = q_f$, and $q_0 = q_i$. The familiar path integral limit is obtained by taking the $N \to \infty$ limit and making the identifications

$$\sum_{k=1}^{N+1} p_k(q_k - q_{k-1}) = \sum_{k=1}^{N+1} p(t_k)\frac{q(t_k) - q(t_k - \epsilon)}{\epsilon}\epsilon \to \int_0^t p(t)\dot{q}(t)dt$$

and

$$\sum_{k=1}^{N+1} H(q_k,p_k)\epsilon \to \int_0^t H(q(t),p(t))dt$$

for the exponential terms and defining

$$\prod_{k=1}^{N} \frac{dp_k\, dq_k}{2\pi\hbar} = \mathscr{D}p\, \mathscr{D}q$$

to be the integration measure. Inserting all of these relations into (4) yields

$$\langle q_f|q_i\rangle(t) = \int \frac{dp_{N+1}}{2\pi\hbar} \int \mathscr{D}p\, \mathscr{D}q \exp\left\{\frac{i}{\hbar}\int_0^t [p(t)\dot{q}(t) - H(q(t),p(t))]dt\right\}, \tag{5}$$

which is the familiar PSPI.[2]

With the formal identifications made above, the exponent in (5) represents the classical action in first-order form for a particle traveling from q_i to q_f. From the integration measure, the PSPI would seem to be a sum over all possible paths in phase space between the initial and final state. However, as has often been noted,[2] this interpretation is completely false. Examining the discrete form of (4) shows the intermediate configurations are exceptionally discontinuous and cannot be equated with classical paths, a failure sometimes referred to as "the problem of unruly paths."

For Hamiltonians at most quadratic in the momenta, in particular those of the form $p^2/2m + V(x)$, the p integrations may be carried out explicitly, either in the discrete form (4) or formally through a functional stationary phase in (5). Taking the former approach yields

$$\begin{aligned}\int \frac{dp_{N+1}}{2\pi\hbar}\left[\prod_{k=1}^{N}\frac{dp_k}{2\pi\hbar}\right]\exp\left\{\frac{i}{\hbar}\sum_{k=1}^{N+1} p_k(q_k - q_{k-1}) - \epsilon\frac{p_k^2}{2m}\right\} &= \int\left[\prod_{k=1}^{N+1}\frac{dp_k}{2\pi\hbar}\right]\exp\left\{\frac{i\epsilon}{2m\hbar}\right. \\ &\quad \left.\times\sum_{k=1}^{N+1}(p_k - m\dot{q}_k)^2 - m^2\dot{q}_k^2\right\} \\ &= \left[\sqrt{\frac{m}{2\pi i\epsilon\hbar}}\right]^{N+1}\exp\left\{\frac{i}{\hbar}\int_0^t \frac{m\dot{q}^2(t)}{2}dt\right\},\end{aligned} \tag{6}$$

where the continuum limit has been taken in the exponential of the last line. Inclusion of the potential terms results in Feyman's original expression

$$\langle q_f|q_i\rangle(t) = \int \mathscr{D}q \exp\left\{\frac{i}{\hbar}\int_0^t L\,dt\right\}. \tag{7}$$

Here, L is the classical action in second-order form

$$L = m\dot{q}^2(t)/2 - V(q) \tag{8}$$

and the integration measure is given by

$$\mathscr{D}q = \sqrt{\frac{m}{2\pi i\epsilon\hbar}}\prod_{k=1}^{N}\sqrt{\frac{m}{2\pi i\epsilon\hbar}}\,dq_k. \tag{9}$$

As stated before the "paths" in this expression are continuous but nowhere differentiable. Problematic though this might be, such paths are far less "unruly" than those of the PSPI; the functional Legendre transformation has greatly smoothed the integrand.

In a similar vein, a path integral expression may be formulated through the insertion of complete sets of coherent states.[9] Although the initial and final states are taken to be coherent states, it is crucial to remember that what makes this a coherent state path integral is the resolution of unity in terms of coherent states inserted an infinite number of times. As before, a transformation function

$$\langle z_f|z_i\rangle(t) = \langle z_f|e^{-i\hat{H}t/\hbar}|z_i\rangle \tag{10}$$

is examined by dividing the time into $N+1$ intervals. Inserting a complete set of coherent states (of the same σ as the initial and final states) N times yields

$$\begin{aligned}\langle z_f|z_i\rangle(t) &= \int\prod_{k=1}^{N}\left[\frac{d^2z_k}{\pi}\right]\langle z_f|e^{-i\hat{H}\epsilon/\hbar}|z_N\rangle \\ &\quad\times\langle z_N|e^{-i\hat{H}\epsilon/\hbar}|z_{N-1}\rangle\cdots\langle z_2|e^{-i\hat{H}\epsilon/\hbar}|z_i\rangle \\ &\quad\times\langle z_1|e^{-i\hat{H}\epsilon/\hbar}|z_i\rangle \\ &= \pi^{-N}\int\left(\prod_{k=1}^{N} d^2z_k\right)\prod_{j=1}^{N+1}\langle z_j|e^{-i\hat{H}\epsilon/\hbar}|z_{j-1}\rangle,\end{aligned} \tag{11}$$

where $z_{N+1} = z_f$ and $z_0 = z_i$. The short time transformation elements may be approximated by

$$\begin{aligned}\langle z_j|e^{-i\hat{H}\epsilon/\hbar}|z_{j-1}\rangle &\approx \langle z_j|(1 - i\epsilon\hat{H}/\hbar)|z_{j-1}\rangle \\ &= \langle z_j|z_{j-1}\rangle(1 - i\epsilon H(z_j^*,z_{j-1})/\hbar) \\ &\approx \langle z_j|z_{j-1}\rangle\exp[(-i\epsilon/\hbar)H(z_j^*,z_{j-1})],\end{aligned} \tag{12}$$

where

$$H(\alpha^*,\beta) = \langle\alpha|\hat{H}|\beta\rangle/\langle\alpha|\beta\rangle$$

is a c-number symbol of the Hamiltonian (commonly called the "ordered symbol" in the language of pseudodifferential operators).[10,11] From the familiar overlap of two coherent states

$$\langle\alpha|\beta\rangle = \exp(\alpha^*\beta - \tfrac{1}{2}|\alpha^*|^2 - \tfrac{1}{2}|\beta|^2),$$

the exponent in (12) may be written

$$\begin{aligned}\sum_{j=1}^{N+1}\left\{z_j^*z_{j-1} - \frac{1}{2}|z_j^*|^2 - \frac{1}{2}|z_{j-1}|^2 - \frac{i\epsilon}{\hbar}H(z_j^*,z_{j-1})\right\} &= \sum_{j=1}^{N+1}\left\{-\frac{1}{2}(z_j^*(z_j - z_{j-1}) - (z_j^* - z_{j-1}^*)z_j)\right. \\ &\quad \left. - \frac{i\epsilon}{\hbar}H(z_j^*,z_{j-1})\right\}.\end{aligned} \tag{13}$$

Multiplying and dividing the first term by ϵ and defining

$$\frac{dz_j}{dt} = \frac{z_j - z_{j-1}}{\epsilon}$$

yields

$$\begin{aligned}\langle z_f|z_i\rangle(t) = \int\prod_{k=1}^{N}\frac{d^2z_k}{\pi} &\times\exp\left\{\sum_{j=1}^{N+1}\epsilon\left[-\frac{1}{2}\left(z_j^*\frac{dz_j}{dt} - \frac{dz_j^*}{dt}z_{j-1}\right)\right.\right. \\ &\left.\left. - \frac{i}{\hbar}H(z_j^*,z_{j-1})\right]\right\}.\end{aligned} \tag{14}$$

To pass to the path integral limit, one commonly replaces z_{j-1} by z_j, which introduces errors of order $\epsilon(z_j - z_{j-1})$. This approximation and the continuum limit result in the conventional expression for the CSPI

$$\langle z_f|z_i\rangle(t) = \int \mathscr{D}z\,\mathscr{D}z^* \exp\left[\frac{i}{\hbar}\int_0^t \left\{\frac{i\hbar}{2}(z^*\dot{z} - \dot{z}^*z) - H(z^*,z)\right\}dt\right], \tag{15}$$

where

$$\mathscr{D}z\,\mathscr{D}z^* = \prod_{i=1}^{N}\frac{d^2z_i}{\pi} = \prod_{i=1}^{N}\frac{dp_i\,dq_i}{2\pi\hbar}, \tag{16}$$

and the exponent may be interpreted as the action in first-order form for the corresponding classical coherent state evolution. The second form of (16) indicates the path integral may be taken over the "phase space" of position and momentum expectation values of the coherent states. Again, the numerical value of H will generally differ from simply replacing the operators $\hat{q}$ and $\hat{p}$ by their respective eigenvalues

$$q_j = (\sigma/\sqrt{2})(z_j + z_j^*)$$

and

$$p_j = (-i\hbar/\sqrt{2}\sigma)(z_j - z_j^*).$$

Expression (15) appears to give a well-defined path integral expression for the time-dependent coherent state overlap; however, because of the discontinuous nature of the phase space paths the neglected terms $\epsilon(z_j - z_{j-1})$ can be of order ϵ and thus could contribute to the action. As Schulman has speculated, "My own guess is that they can contribute and that this contribution will be related to the operator ordering problem in quantum mechanics." [2] The veracity of this prediction will be demonstrated shortly.

III. LEGENDRE TRANSFORMATION OF THE COHERENT STATE PATH INTEGRAL

As stated above, the CSPI represents the most viable candidate for a true functional integral formulation of the quantum propagator. However, the canonical form (15) has been shown[2,6] to be plagued by many of the same problems which bedevil the PSPI; in particular, the phase space paths prove highly unruly. In addition, Schulman's speculation about the $\epsilon(z_j - z_{j-1})$ error terms, if correct, would imply that (15) is not necessarily complete through order ϵ, and thus an ill conceived quantity at best. Clearly, any attempt to elevate the CSPI to a true functional integral must resolve both of these issues.

Attacking the problem of the unruly paths first, the discontinuous nature of the phase space paths arises from the fact that while the exponent in (15) may be interpreted as a classical action in first-order form, the measure (16) contains $2N$ integration variables corresponding to the N phase space points in the action. Since a single point in phase space suffices to specify the classical path, straightforward interpretation of (15) as a sum over phase space paths proves impossible.

In the similar case of the PSPI, arguments have been given for a "generalized" interpretation of phase space paths[2] with the time divided into $2N$ intervals. Rather than debate the merits of such a loosened interpretation, the most satisfactory resolution is attained by integrating over the momenta, yielding a classical action in second-order form with N integration variables in the functional measure, the familiar FPI. Thus one might reason, in analogy with the PSPI to FPI transformation, that a functional Legendre transformation of the coherent state action from first-order to second-order form would result in paths, like those of the FPI, which are continuous but not differentiable. For Hamiltonians of the form $H = p^2/2m + V(x)$ this Legendre transformation can be carried out by simply performing the Gaussian momentum integrals in (15).

The $\epsilon(z_j - z_{j-1})$ terms prove more subtle. These errors arise from two separate sources, the approximation to the "primitive action"

$$\epsilon\left(z_j^*\frac{dz_j}{dt} - \frac{dz_j^*}{dt}z_{j-1}\right) \approx \epsilon(z_j^*\dot{z}_j - \dot{z}_j^* - z_j) \tag{17}$$

and the approximation to the ordered Hamiltonian symbol

$$\epsilon H(z_j^*,z_{j-1}) \approx \epsilon H(z_j^*,z_j) + \frac{\partial H}{\partial z}\epsilon(z_j - z_{j-1}) + \cdots. \tag{18}$$

Accepting for the moment the approximation to the primitive action (17), it is generally possible to eliminate the $\epsilon(z_j - z_{j-1})$ error terms in (18) by developing a slightly different version of the CSPI based on a different classical symbol for the Hamiltonian.[9]

Again, consider the coherent state transformation function (10) with the time divided into N parts, assumed small. Then,

$$\langle z_f|z_i\rangle(t) = \langle z_f|[\exp(-i\epsilon\hat{H}/\hbar)]^N|z_i\rangle. \tag{19}$$

Rather than insert $N-1$ resolutions of the identity, it proves extremely convenient at this point to take advantage of a particular aspect of the coherent state basis; specifically, because the coherent states form an overcomplete set it is possible to represent an operator, $\hat{T}$, by its diagonal elements alone,

$$\hat{T} = \int |z\rangle\Phi_T(z^*,z)\langle z|\frac{d^2z}{\pi}, \tag{20}$$

where $\Phi_T(z^*,z)$ is a scalar function, which I call the Glauber function for the operator $\hat{T}$ (Ref. 12). Like the ordered symbol, H, the Glauber function may be taken as a c-number representation of an operator, referred to as the antiordered symbol.[11] As the names imply, the differences between the ordered and antiordered symbols correspond to different choices of operator ordering. For example, for the operator $\hat{x}^2$, the ordered symbol is given by $q^2 + \sigma^2/2$ while the antiordered symbol is $q^2 - \sigma^2/2$ with the differences between the two symbols arising from commutators of $\hat{a}$ and $\hat{a}^\dagger$. For the unfamiliar, a basic description of the derivation and properties of the Glauber function are provided in the Appendix; in particular, it is shown that for operators which are a function of $\hat{x}$ or $\hat{p}$ only the Glauber function depends only on q or p, respectively, a fact which will be used extensively below.

Via use of the Glauber representation, one may write

$$e^{-i\epsilon\hat{H}/\hbar} \approx \int |z\rangle \Phi_H(z^*,z) \langle z| \frac{d^2z}{\pi} \tag{21}$$

where

$$\Phi_H(z^*,z) = e^{-i\epsilon h(z^*,z)/\hbar}, \tag{22}$$

$h(z^*,z)$ is the Glauber function for the Hamiltonian operator

$$\hat{H} = H(\hat{p},\hat{x}) = \int |z\rangle h(z^*,z) \langle z| \frac{d^2z}{\pi}, \tag{23}$$

and (21) holds to $O(\epsilon)$. Use of (21) N times in (19) yields

$$\langle z_f|z_i\rangle(t) = \pi^{-N} \int \prod_{k=1}^{N} d^2z_k \times \exp\left\{ \sum_{j=1}^{N+1} \epsilon \left[-\frac{1}{2}\left(z_j^* \frac{dz_j}{dt} - \frac{dz_j^*}{dt} z_{j-1} \right) - \frac{i}{\hbar} h(z_j^*, z_j) \right] \right\}. \tag{24}$$

The path integral limit is attained by again replacing z_{j-1} with z_j in the primitive action only and taking the limit $N \to \infty$:

$$\langle z_f|z_i\rangle(t) = \int \mathcal{D}z\, \mathcal{D}z^* \exp\left[\frac{i}{\hbar} \int_0^t \left\{ \frac{i\hbar}{2}(z^*\dot{z} - \dot{z}^*z) - h(z^*,z) \right\} dt \right], \tag{25}$$

with the integration measure the same as in (16). Once again, the exponent in (25) may be interpreted as a classical action in first-order form with the integration over the phase space of the position and momentum expectation values of the coherent states.

At this point Schulman's speculation has been proven; the terms $\epsilon(z_j - z_{j-1})$ do contribute to $O(\epsilon)$ in that $H(z^*,z) \neq h(z^*,z)$ so the coherent state actions of (15) and (25) differ. Moreover, this difference stems from different prescriptions for operator ordering (ordered versus antiordered) meaning the two actions agree classically but have different quantum corrections.[10]

For Hamiltonians of the form $\hat{p}^2 + V(\hat{x})$ the Glauber function may be written in the simple form

$$h(p,q) = p^2/2m - \hbar^2/2\sigma^2 + v(q), \tag{26}$$

where $v(q)$ is the Glauber function for the operator $V(\hat{x})$ and does not depend on p. From (23), it is apparent the use of the Glauber representation has both eliminated the second source of $\epsilon(z_j - z_{j-1})$ errors and explicitly separated the position and momentum variables in the coherent state action. The only dependence on the momenta will be an N-dimensional (complex) Gaussian which can be performed exactly either in discrete form or by functional steepest descents. Taking the second approach, one finds

$$\int_0^t \left\{ \frac{i\hbar}{2}(z^*\dot{z} - \dot{z}^*z) \right\} dt = \frac{1}{2}\int_0^t (\dot{q}p - \dot{p}q)dt = -\frac{q_fp_f - q_ip_i}{2} + \int_0^t p\dot{q}\, dt, \tag{27}$$

so that the steepest descent path, with $h(p,q)$ as in (26), for the p integration is given by

$$\dot{q} = \frac{\partial h}{\partial p} = \frac{p}{m}, \tag{28}$$

which is the familiar condition for the Legendre transform of the action from first- to second-order form. Carrying out the functional integration finally yields

$$\langle z_f|z_i\rangle(t) = \int \mathcal{D}q \exp\left[\frac{i}{\hbar} \int_0^t \mathcal{L}\, dt \right], \tag{29}$$

where

$$\mathcal{L}(q,\dot{q}) = m\dot{q}^2 - h(q,\dot{q}), \tag{30}$$

and the measure is given by

$$\mathcal{D}q = e^{-i(q_fp_f - q_ip_i)/2\hbar} \prod_{j=1}^{N} \sqrt{\frac{m}{2\pi i\epsilon\hbar}}\, dq_j. \tag{31}$$

Comparison of (29)–(31) with the FPI expressions (7)–(9) shows both striking similarities and a few significant differences. First, although the integrands are superficially identical, the "Lagrangians" $L(q,\dot{q})$ and $\mathcal{L}(q,\dot{q})$ will differ. Here, L will be the classical Lagrange function $T - V$ while $\mathcal{L}$ will be the classical Lagrangian plus corrections of higher order in $\hbar$. Thus the equations of motion derived from the actions in (7) and (29) will differ as well, but will coincide in the $\hbar \to 0$ limit, a reflection of the Ehrenfest relations[13] regarding the evolution of a coordinate eigenstate versus a wavepacket. The measures (9) and (31) differ by an irrelevant constant phase $(q_fp_f - q_ip_i)/2\hbar$ which may be absorbed into the definition of the coherent states, and one factor of $\sqrt{m/2\pi i\epsilon\hbar}$ which arises from the differing normalizations of coherent and coordinate states [that is, as $t \to 0$ (10) becomes unity while (1) is a delta function].

A far more subtle difference between (29) and the FPI lies in the boundary conditions on the action functional. In (29) both the initial position and velocity are specified, as well as the final position and velocity. Thus the "paths" for (29) cannot be interpreted as classical in the conventional sense since the "Lagrangian" leads to a classical mechanics with overspecified boundary conditions. In general, the initial state with classical variables $(q_i,\dot{q}_i)$ will not evolve in time t to $(q_f,\dot{q}_f)$ so there will be no point of stationary phase in the space of real q. Nonetheless, the quantum amplitude for the transition cannot be zero (although for coherent states it generally will be exponentially small) so the asymptotic $\hbar \to 0$ approximation to (29) generates an analytic continuation of classical mechanics to processes not allowed in conventional Lagrangian dynamics. Interesting though these issues might be, they lie outside the main thrust of this analysis; thus, the stationary paths of (29) and their extension into the fixed energy domain will be examined separately in a forthcoming paper.[14]

Although (29) may be interpreted as a sum over continuous but nondifferentiable paths in the manner of the FPI, the goal of a true functional integral remains unattained, for both (29) and the FPI are beset with the same complex and infinite pseudomeasure for integration over real intermediate positions. Moreover, and most surprisingly, (29), like (7), has an integrand of unit magnitude, meaning all possible paths contribute to the integral with equal weight. This violates the expectation, noted above, that only a fraction of the available phase space would contribute significantly to the quantum evo-

lution. If, for example, (29) described a coherent state evolving in a harmonic potential, then the Gaussian wavepacket would remain localized in position and momentum and any paths in phase space away from the classical one would be exponentially small. This seeming anomaly cannot be simply argued away as a result of integrating over the coherent state momenta since, in general, performing the p integral in the coherent state resolution of unity is not equivalent to inserting the coordinate resolution. Either the hope that the CSPI could lead to a true functional integral is wrong or else (29) itself is somehow incomplete to $O(\epsilon)$, with the only possible source of such an error being the $\epsilon(z_j - z_{j-1})$ term neglected in the primitive action.

IV. FINAL CORRECTIONS TO THE CSPI

Returning to the discrete form of the CSPI (24), it is possible to perform the momentum integrals analytically without taking $z_j \approx z_{j-1}$ for Hamiltonians quadratic in the momenta. Moreover, the integrations can be carried out prior to taking the path integral limit, i.e., without interchanging the $N \to \infty$ limit with the order of integration. Since the approximation to the primitive action is now the object of interest, it proves simplest initially to consider the nondynamical case, i.e., $\hat{H} = 0$. Then the overlap of two coherent states is constant in time and may be written

$$\begin{aligned}\langle z_f | z_i \rangle &= \exp\left[z_f^* z_i - \tfrac{1}{2}\left(|z_f^*|^2 + |z_i|\right)\right] \\ &= \int \prod_{i=1}^{N} \left[\frac{dp_i\, dq_i}{2\pi\hbar}\right] \exp\left\{ -\frac{\sigma^2}{\hbar^2}\mathbf{p}A\mathbf{p} - \frac{1}{\sigma^2}\mathbf{q}B\mathbf{q} \right. \\ &\quad \left. + \frac{i}{\hbar}\mathbf{p}C\mathbf{q} + \mathbf{p}\cdot\boldsymbol{\alpha} + \mathbf{q}\cdot\boldsymbol{\beta} - \frac{|z_f^*|^2 + |z_i|^2}{2}\right\}, \end{aligned} \tag{32}$$

where $\mathbf{q}$ and $\mathbf{p}$ are the N-dimensional vectors of intermediate integration variables,

$$\boldsymbol{\alpha} = \frac{i\sigma}{\sqrt{2}\hbar}\begin{pmatrix} -z_i \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \\ z_f^* \end{pmatrix}, \quad \boldsymbol{\beta} = \frac{1}{\sqrt{2}\sigma}\begin{pmatrix} z_i \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \\ z_f^* \end{pmatrix}, \tag{33}$$

and the matrices A and B are identical and given by

$$A = B = \begin{pmatrix} \frac{1}{2} & -\frac{1}{4} & 0 & 0 & \cdots & & \\ -\frac{1}{4} & \frac{1}{2} & -\frac{1}{4} & 0 & 0 & \cdots & \\ 0 & -\frac{1}{4} & \frac{1}{2} & -\frac{1}{4} & 0 & & \\ 0 & 0 & \ddots & \ddots & \ddots & & \\ \vdots & & & & & & \\ & & & & \ddots & \ddots & \ddots \\ & & & & & -\frac{1}{4} & \frac{1}{2} & -\frac{1}{4} \\ & & & & & 0 & -\frac{1}{4} & \frac{1}{2} \end{pmatrix}.$$

The matrix C is a "difference matrix" that generates terms corresponding to velocities and is given by

$$C = \begin{pmatrix} 0 & \frac{1}{2} & 0 & 0 & \cdots & & \\ -\frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 & \cdots & \\ 0 & -\frac{1}{2} & 0 & \frac{1}{2} & 0 & & \\ 0 & 0 & \ddots & \ddots & \ddots & & \\ \vdots & & & & & & \\ & & & & \ddots & \ddots & \ddots \\ & & & & & -\frac{1}{2} & 0 & \frac{1}{2} \\ & & & & & 0 & -\frac{1}{2} & 0 \end{pmatrix}.$$

For convenience, the dimensional constants σ and $\hbar$ will be temporarily scaled to one. They can easily be recovered by a simple dimensional analysis at any intermediate point.

Before proceeding, it is worth noting that via straightforward definitions and retaining all terms of $O(\epsilon)$, (33) may be expressed in the path integral limit[15] as

$$\langle z_f | z_i \rangle = \int \mathscr{D}p\, \mathscr{D}q \exp\left[i \int_0^t \left(\frac{1}{2}(p\dot{q} - \dot{p}q) + \frac{1}{4} i\epsilon(\dot{p}^2 + \dot{q}^2)\right) dt' \right]. \tag{34}$$

Here, ϵ corresponds to the time step and is strictly zero as the number of integrations becomes infinite. As demonstrated by Klauder, ϵ plays the part of the smile of the Cheshire cat, that is, the last remaining sign of the discrete time slicing prescription as $N \to \infty$. Nonetheless, only by keeping this "smile" can the correct (discontinous) boundary conditions on the CSPI be derived.[15] In terms of finding a true functional integral, (34) represents a dramatic improvement. As mentioned, Klauder and Daubechies[6] have shown (34) may be interpreted as an example of "well-defined phase space path integrals involving Wiener measure on continuous phase-space paths in the limit that the diffusion constant diverges."

Nonetheless, (34), while formally an improvement, remains problematic from a practical point of view. There are again $2N$ integration variables corresponding to N points in the phase space of paths making these "paths" again "unruly" [indeed, according to Klauder, it was the desire to include such unruly paths which originally motivated (34) (Ref. 15)]. In fact, the paths in this expression are in some senses even more unruly since the action inherently depends not just upon p but $\dot{p}$ as well. As indicated by the previous examples, one can ameliorate this difficulty by performing a Legendre transformation to the variables q and $\dot{q}$. What proves tantalizing about (34) is that, although the integrand is of unit modulus in the limit as

$\epsilon \to 0$, for finite N the integrand contains a clear maximum with the "smile" providing Gaussian damping. As shall be demonstrated below, the "smile" will ultimately survive the Legendre transformation, yielding a path integral expression with both continuous paths and Wiener measure for finite N.

To this end, we turn our eyes to Eq. (33) and perform the N momentum integrals. From standard rules of multidimensional Gaussian integration, this yields

$$\langle z_f|z_i\rangle = \frac{e^{-(|z_f^*|^2+|z_i|^2)/2}}{\sqrt{\mathrm{Det}_N A}} \int \frac{\Pi_i\, dq_i}{(4\pi)^{N/2}} \times \exp\left[-\mathbf{q}\left(A + C^T \frac{A^{-1}}{4} C\right)\mathbf{q} + \mathbf{q}\cdot\boldsymbol{\beta} + i\boldsymbol{\alpha}\frac{A^{-1}}{4} C\mathbf{q} + i\mathbf{q}C^T\frac{A^{-1}}{4}\boldsymbol{\alpha} + \boldsymbol{\alpha}\frac{A^{-1}}{4}\boldsymbol{\alpha}\right], \tag{35}$$

where the matrix A^{-1} must be symmetric. Considered as a function of q, the properties of the integrand will be dominated by the terms quadratic in q and thus one must know the form of A^{-1}. Tridiagonal matrices such as A and C are known as Jacobi matrices.[2] All of the necessary properties of Jacobi matrices can be derived fairly straightforwardly; the results will be outlined below.

For a generic $N \times N$ Jacobi matrix of the form

$$J = \begin{pmatrix} a & b & 0 & 0 & \cdots & & \\ b & a & b & 0 & 0 & \cdots & \\ 0 & b & a & b & 0 & & \\ 0 & 0 & \ddots & \ddots & \ddots & & \\ \vdots & & & & & & \\ & & & & \ddots & \ddots & \ddots \\ & & & & & b & a & b \\ & & & & & 0 & b & a \end{pmatrix},$$

a recurrence relationship for the determinate can be found by expanding in minors

$$D_N = aD_{N-1} - b^2 D_{N-2} \tag{36}$$

yielding a second-order finite difference equation with constant coefficients. Solutions to (36) can be written as

$$D_N = c_1 y_1^N + c_2 y_2^N, \tag{37}$$

where y_1 and y_2 are the two roots of

$$y^2 - ay + b^2 = 0 \tag{38}$$

and the coefficients c_1 and c_2 are found by imposing

$$D_0 = 1 \quad \text{and} \quad D_1 = a. \tag{39}$$

As can be directly verified, the inverse of the Jacobi matrix is given by

$$(J_N^{-1})_{jk} = (D_{j-1} D_{N-k}/D_N)(-b)^{k-j}, \quad j \leqslant k, \tag{40}$$

with the $j > k$ elements given (since J^{-1} is symmetric) by interchanging j and k.

For the case at hand, $a = 1/2$ and $b^2 = 1/16$ making $y = 1/4$ a double root of (38). Thus

$$D_N = c_1/4^N + Nc_2/4^N,$$

and conditions (39) give $c_1 = c_2 = 1$. Hence,

$$D_N = (1+N)/4^N, \tag{41}$$

and the elements of the inverse matrix will be given by

$$A_{jk}^{-1}/4 = j[1 - k/(N+1)], \quad j \leqslant k$$
$$= k[1 - j/(N+1)], \quad j \geqslant k. \tag{42}$$

From the form of the C, one finds

$$[C^T(A^{-1}/4)C]_{ij} = \tfrac{1}{16}\{A^{-1}_{i+1,j+1} - A^{-1}_{i-1,j+1} - A^{-1}_{i+1,j-1} + A^{-1}_{i-1,j-1}\}. \tag{43}$$

Taking $i \leqslant j$ for definiteness, relation (42) may be used directly in this expression for all but the third term in which, for $i = j$ or $i = j-1$, the indices must be reversed. Additionally, the cases $i = 1$ and $j = N$ must be handled separately. Examination of all these cases yields

$$C^T\frac{A^{-1}}{4}C = \begin{pmatrix} \frac{1}{2} & \frac{1}{4} & 0 & 0 & \cdots & & \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} & 0 & 0 & \cdots & \\ 0 & \frac{1}{4} & \frac{1}{2} & \frac{1}{4} & 0 & & \\ 0 & 0 & \ddots & \ddots & \ddots & & \\ \vdots & & & & & & \\ & & & & \ddots & \ddots & \ddots \\ & & & & & \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ & & & & & 0 & \frac{1}{4} & \frac{1}{2} \end{pmatrix} - [1/(N+1)]S, \tag{44}$$

where the matrix S has every element equal to one, $S_{ij} = 1$, $i,j = 1 \ldots N$. Thus (35) may be written

$$\langle z_f|z_i\rangle = \frac{e^{-(|z_f^*|^2+|z_i|^2)/2}}{\sqrt{\mathrm{Det}_N A}} \int \frac{\Pi_i\, dq_i}{(4\pi)^{N/2}} \times \exp\left[-\mathbf{q}(I - \epsilon S)\mathbf{q} + \mathbf{q}\cdot\boldsymbol{\beta} + i\boldsymbol{\alpha}\frac{A^{-1}}{4} C_{\mathbf{q}} + i\mathbf{q}C^T\frac{A^{-1}}{4}\boldsymbol{\alpha} + \boldsymbol{\alpha}\frac{A^{-1}}{4}\boldsymbol{\alpha}\right], \tag{45}$$

where $\epsilon = 1/(N+1)$ and I is the $N \times N$ identity matrix. Recognizing the third and fourth terms are identical, and performing an analysis similar to the above on the matrix product CA^{-1}, one arrives at

$$\langle z_f|z_i\rangle = \frac{e^{-(|z_f^*|^2+|z_i|^2)/2}}{\sqrt{\mathrm{Det}_N A}} \int \frac{\Pi_i\, dq_i}{(4\pi)^{N/2}} \times \exp\left[-\mathbf{q}\left(I - \frac{1}{N+1}S\right)\mathbf{q} - \frac{1}{2}z_i^2\left(\frac{N}{N+1}\right) - \frac{1}{2}z_f^{*2}\left(\frac{N}{N+1}\right) + \frac{z_f^* z_i}{N+1} + \sqrt{2}(z_f^* + z_i)\sum_{j=1}^{N}\frac{q_j}{N+1}\right]. \tag{46}$$

Performing this integral, and understanding its behavior, requires knowledge of the inverse and determinate of $I - \epsilon S$. The determinate may be written as

$$\begin{aligned}\mathrm{Det}(I-\epsilon S) &= \epsilon_{ijkl\ldots}(I-\epsilon S)^1_i(I-\epsilon S)^2_j(I-\epsilon S)^3_k\cdots \\ &= \epsilon_{ijkl\ldots}I^1_iI^2_jI^3_k\cdots - \epsilon\{\epsilon_{ijkl\ldots}S^1_iI^2_jI^3_k\cdots \\ &\quad + \epsilon_{ijkl\ldots}I^1_iS^2_jI^3_k\cdots + \cdots\} \\ &\quad + \text{higher orders in } \epsilon,\end{aligned}$$

which is just a sum of determinants. But all terms of higher order in ϵ contain at least two rows from S and hence have determinate zero; moreover, all terms of $O(\epsilon)$ are the same so

$$\begin{aligned}\mathrm{Det}_N(I-\epsilon S) &= \mathrm{Det}_N I - N\epsilon\,\mathrm{Det}_N S' \\ &= 1 - N\epsilon = 1/(N+1),\end{aligned} \tag{47}$$

where S' is the $N\times N$ identity matrix with one row (or column) replaced by a row (or column) of S. It is straightforward to show the eigenvalues of $I-\epsilon S$ are

$$\lambda_i = \begin{cases} 1 & i=1 \text{ to } N-1; \\ 1/(N+1) & i=N, \end{cases} \tag{48}$$

and the normalized eigenvector corresponding to λ_N is easily found to be

$$v_N = \frac{1}{\sqrt{N}}\begin{pmatrix}1\\1\\1\\\vdots\\1\end{pmatrix}, \tag{49}$$

which represents a "breathing" mode, that is, a uniform displacement in all the coordinates. The inverse of $I-\epsilon S$ can be found by considering the cofactor matrix and expanding in minors and is given by $I+S$, as can be directly verified. Using the expression for the inverse and determinate of A and $I-\epsilon S$ and the rules of Gaussian integration, it is straightforward to show (46) indeed gives the correct expression for the overlap of two coherent states. Moreover, by defining $q_0 = z_i/\sqrt{2}$ and $q_{N+1} = z_f^*/\sqrt{2}$ expression (46) can be written in the simple form

$$\langle z_f|z_i\rangle = \int \mathscr{D}q\exp\left[-\mathbf{q}(I-\epsilon S)\mathbf{q} - \frac{1}{2}(|z_f^*|^2+|z_i|^2)\right] \tag{50}$$

where q, I, and S all now have indices running from 0 to $N+1$. In the $N\to\infty$ limit, the exponent in (50) does not become an expression of classical action, but the integral has other intriguing properties. Consider that (46) can alternatively be written

$$\langle z_f|z_i\rangle \sim \int \mathscr{D}q\exp\left[-\mathbf{q}(I-\epsilon S)\mathbf{q} + \sqrt{2}\epsilon(z_f^*+z_i)\mathbf{q}\cdot\mathbf{L}\right], \tag{51}$$

with the vector L given by

$$L = \begin{pmatrix}1\\1\\1\\\vdots\\1\end{pmatrix}.$$

The integrand of this multidimensional Gaussian integral has a maximum at

$$\begin{aligned}q_j &= \frac{(z_f^*+z_i)}{\sqrt{2}}\epsilon\sum_{k=1}^{N}(I+S)_{jk}L_k \\ &= (z_f^*+z_i)/\sqrt{2},\end{aligned} \tag{52}$$

for all $j=1\ldots N$. For integration variables restricted to the real axis the integrand of (51) will have a maximum at $q_j = (q_f+q_i)/2$ and Gaussian damping away from that point. Hence, (51) is an ideal candidate for evaluation through Monte Carlo integration techniques.

Expression (50), like (34), has arisen from inclusion of all terms $O(\epsilon)$. However, in (34) the order of integration and the $N\to\infty$ limit have been interchanged while, in (50), the momentum integrations are performed before the path integral limit, yielding a representation for the overlap with distinctly different properties. However, the exponent in (50), unlike that in (34), does not appear to correspond to a classical action in the path integral limit. One could reasonably ask whether the $\hat{H}=0$ case might not be pathological, i.e., that including dynamics would eliminate the damping factors in the path integral integrand and/or make the exponent of (50) correspond to a classical action (certainly in the FPI, there exists no "action" to put in the $\hat{H}=0$ path integral since the overlap of two coordinate states reduces to an infinite product of Dirac δ functions).

Nonetheless, an analysis similar to the above proves there is nothing unusual about a zero Hamiltonian. Using the Glauber representation for $\hat{H}=\hat{p}^2/2m+V(\hat{x})$ and formulating the time-dependent overlap will, as noted above, add a sum of terms to the exponent (32) of the form $h(p_j,q_j)=p_j^2/2m+v(q_j)$. This simply changes the diagonal elements of the matrix A to $(1-i\epsilon)/2$, where $\epsilon=t/N$. Inclusion of these terms splits the degeneracy in the roots of the finite difference equation (36). Proceeding as before, one finds

$$D_n = c_+y_+^n + C_-y_-^n,$$

where

$$y_\pm = \tfrac{1}{4}(1+i\epsilon\pm\sqrt{2i\epsilon-\epsilon^2})$$

and the conditions $D_0=1$, $D_1=(1+i\epsilon)/2$ yield

$$c_\pm = \pm\left[2/(\sqrt{2i\epsilon-\epsilon^2})y_\pm\right].$$

Anticipating the path integral limit, D_n can be expanded to first order in ϵ yielding

$$D_n \approx [(n+1)/4^n](1+in\epsilon)$$

and the elements of A^{-1} are given to first order in ϵ by

$$A^{-1}_{jk} = 4^j(1-k/(N+1))\{1-i\epsilon(k-j+1)\},\quad j\leqslant k, \tag{53}$$

with the $j>k$ elements given by the symmetry of A^{-1}. By separating the real and imaginary parts of A^{-1}, the time-dependent overlap of two coherent states can be written in the form

$$\langle z_f|z_i\rangle(t) = \mathscr{N}\int \mathscr{D}q\rho(\mathbf{q})\exp\left[\frac{i}{\hbar}F(\mathbf{q})\right], \tag{54}$$

where $\mathscr{N}$ is a normalization factor, $\rho(\mathbf{q})$ is an N-dimensional Gaussian probability density again given by

$$\rho(\mathbf{q}) = \exp[-\mathbf{q}(I-S/(N+1))\mathbf{q}] \tag{55}$$

and

$$F(\mathbf{q}) = -i\hbar\mathbf{q}\cdot\boldsymbol{\beta} + \alpha\frac{A^{-1}}{4}C\mathbf{q}$$

$$+ \epsilon\mathbf{q}C^{T}MC\mathbf{q} - \epsilon\sum_{j=1}^{N} v(q_j). \tag{56}$$

Here the matrix M can be thought of as a mass matrix (since $Cq \sim \dot{q}$) and arises from the imaginary part of A^{-1} with elements given by

$$M_{jk} = j(k+1-j)(1-k/(N+1)), \quad j \leqslant k,$$

where, once again, the $j > k$ elements are given by the symmetry of the matrix. Note that expression (54), like (50), has Wiener (Gaussian damped) measure for finite N but, from the expression for the mass matrix, the exponent of (54) cannot be interpreted in the path integral limit as a classical action.

V. DISCUSSION

The results of the preceding section have demonstrated the delicacy of the path integral limit for the CSPI. Performing the momentum integrals before passing to this limit yields an integral expression which, while of well-defined measure, cannot be straightforwardly interpreted as a sum over classical paths. The problem of unruly paths has been eliminated, but while the "paths" are continuous the mass matrix couples all the discrete velocities to one another in a nontrivial way with no obvious limit as a Riemann integral.

Nonetheless, although (54) seems to fall short of being a true functional integral, it represents a clear step forward in that the unruly discontinuous paths make an exponentially small contribution to the integrand. Consider the eigenvalues of the matrix $I - \epsilon S$ given in (48). The overlap of two coherent states falls off exponentially with the distance of separation in phase space. Hence, for finite N, a path constructed in phase space from an initial to a final coherent state cannot make an arbitrary excursion without developing an exponentially small contribution to the amplitude; the centers of the coherent states in the path must remain relatively near each other. In this vein, λ_N corresponds to the possible excursion of the path (an interpretation verified by its "breathing mode" eigenvector) while the $\lambda_{1\ldots N-1}$ represent the "tension in the string" connecting the initial and final states. As $N \to \infty$ the string can be arbitrarily long, making arbitrary excursions in phase space as signified by $\lambda_N \to 0$. However, all the other λ_i are constant, so the string still cannot be "stretched" without the integrand becoming exponentially small, Fig. 1. Thus (54) preferentially weights continuous paths from $|z_i\rangle$ to $\langle z_f|$. It is possible that through a judicious coordinate transformation (54) can be expressed, in the $N \to \infty$ limit, as a well-defined functional integral.

In addition, (54) proves extremely useful from a practical point of view. Because the integrand factors into a well-defined probability distribution and an oscillatory part, numerical integration by Monte Carlo techniques[16] proves quite straightforward in comparison to a similar efforts utilizing the FPI[17-19] (which, for real time, contains no probability density, just the oscillatory term). The presence of this Gaussian probability density reflects the localized nature of the coherent states as they evolve in time, making "paths" far from the classical one contribute negligibly to the integrand. Rather than extend an already long paper with this tangential subject, a single simple example will be presented with the general development delineated in a separate publication.[20]

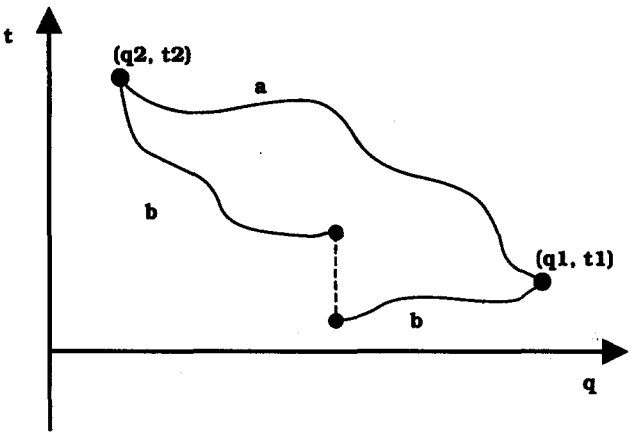

FIG. 1. Two possible paths leading from $(q1,t1)$ to $(q2,t2)$. Path (a) is smooth and contributes appreciably to the integrand of the CSPI, while the contribution of path (b), which has a discontinuity at an intermediate time, is exponentially small.

Figure 2 represents results for the Monte Carlo evaluation of two periods of the time-dependent amplitude $\langle z|e^{-i\hat{H}t/\hbar}|z\rangle$ for the ground state of the harmonic oscillator; that is, $\hat{H} = p^2/2m + m\omega^2x^2/2$, where the convenient choices of $\omega = m = \hbar = 1$ have been taken and $z = 0$. For the ground state, the amplitude is given analytically by $\langle 0|0\rangle(t) = e^{-it/2}$. The calculation was performed using $N = 8$ with 3000 points per integration dimension, and required, literally, but a few

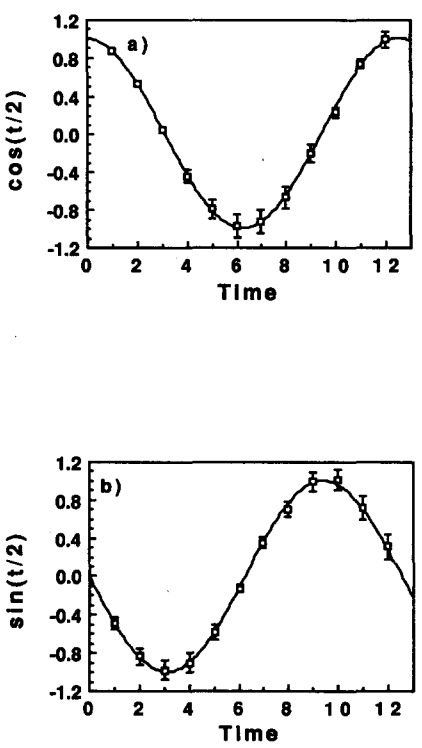

FIG. 2. Real and imaginary parts of the autocorrelation function for the ground state of the harmonic oscillator exactly and by Monte Carlo sampling of the discrete coherent state path integral with $N = 8$. Monte Carlo calculations were performed using 24 000 points, or 3000 points per integration dimension.

seconds on the Ohio Supercomputer Center Cray YMP. Extensions to anharmonic systems in one and two spatial dimensions have proven quite promising.

Finally, lest any confusion arise, it must be noted that the well-established expression for the CSPI, (15), is not in any sense "incorrect." In repeated applications, (15) has yielded perfect results, in particular for the semiclassical approximation and quadratic Hamiltonians, and has been shown completely equivalent to other semiclassical approximations.[21,22] This paper presents "corrections" to the interpretation of (15) as a path integral formulated from the repeated insertion of the coherent state resolution of the identity. Consider the transition amplitude

$$\langle z_f|e^{-i\hat{H}t/\hbar}|z_i\rangle = \int \langle z_f|q_N\rangle\langle q_N|e^{-i\hat{H}t/\hbar}|q_0\rangle\langle p_0|q_0\rangle \times\langle p_0|z_i\rangle dp_0\, dq_0\, dq_N, \tag{57}$$

where three resolutions of the identity in terms of momentum and position eigenstates have been made. Using the PSPI representation of $\langle q_N|q_0\rangle(t)$ in (57) yields an expression identical to (15) except for Gaussian factors in the integrations over p_N and q_0, respectively. In the $N\to\infty$ limit, these factors are irrelevant (formally speaking, they are of measure zero in the functional integration) so (15) is recognized as a PSPI between coherent states. The Hamiltonian symbols that arise from (15) and (57) will, of course, differ, but the differences again correspond to choices of operator ordering; to first order in $\hbar$ the symbols coincide. Stated another way, neglecting the terms of order $\epsilon(z_j - z_{j-1})$ in passing from (13) to (15) is equivalent to approximating the coherent state $|z\rangle$ by $|q\rangle\otimes|p\rangle$. Thus, once again, Schulman's speculation has been shown correct; terms of order $\epsilon(z_j - z_{j-1})$ do contribute to the CSPI, and the contributions are related to questions of operator ordering.

ACKNOWLEDGMENTS

I would like to thank Professor Thomas Beck of the University of Cincinnati, Professors Robert Carlitz and Rob Coalson of the University of Pittsburgh, and Patrick O'Connor of the University of Washington for many helpful and enlightening conversations regarding this work. In addition, Dr. Beck and the Chemistry Department of the University of Cincinnati provided financial support during the writing of this paper. Finally, this work is greatly indebted to the developments in a similar vein of J. R. Klauder and, of course, L. S. Schulman.

Monte Carlo calculations were carried out through a grant from the Ohio State Supercomputer Center.

APPENDIX: THE GLAUBER FUNCTION

In an overcomplete basis, specifying all the matrix elements of an operator is redundant as not all of the elements are independent, i.e., in theory some subset of the matrix elements suffices to specify the operator entirely. In fact, for a wide range of operators—specifically, for the bounded operators commonly encountered in quantum dynamics—it can be shown that the diagonal elements alone define the operator uniquely. That is, an arbitrary operator $\hat{A}$ can be represented as

$$\hat{A} = \int |z\rangle\Phi(z)\langle z|\frac{d^2z}{\pi} \tag{A1}$$

where $\Phi(z)$ is a scalar function of z, called the Glauber function, which weights the contribution of the diagonal elements. Hence, knowledge of Φ and relation (A1) specify $\hat{A}$ completely.

The Glauber function for a given operator may be found as follows: Assume (A1) is correct and calculate the diagonal elements of $\hat{A}$. Then

$$A(z') \equiv \langle z'|\hat{A}|z'\rangle = \int |\langle z'|z\rangle|^2\Phi(z)\frac{d^2z}{\pi}. \tag{A2}$$

From

$$|\langle z'|z\rangle|^2 = \exp\{-|z-z'|^2\},$$

the right-hand side of (A2) is seen to be a convolution, a quantity best handled by Fourier transform. I use the Fourier transform conventions:

$$A(z) = \int e^{-i(kp-\omega q)/\hbar}\tilde{A}(k,\omega)\frac{dk\,d\omega}{2\pi\hbar}, \tag{A3}$$

$$\tilde{A}(k,\omega) = \int e^{+i(kp-\omega q)/\hbar}A(p,q)\frac{dp\,dq}{2\pi\hbar}, \tag{A4}$$

$$\Phi(z) = \int e^{-i(kp-\omega q)/\hbar}\tilde{\Phi}(k,\omega)\frac{dk\,d\omega}{2\pi\hbar}, \tag{A5}$$

$$\tilde{\Phi}(k,\omega) = \int e^{+i(kp-\omega q)/\hbar}\Phi(p,q)\frac{dp\,dq}{2\pi\hbar}. \tag{A6}$$

Inserting (A5) into the right-hand side of (A2) and using

$$z = \frac{1}{\sqrt{2}}\left(\frac{q}{\sigma} + \frac{i\sigma}{\hbar}p\right)$$

then yields

$$A(z') = \int \frac{dk\,d\omega}{2\pi\hbar}\frac{dp\,dq}{2\pi\hbar} \times\exp\left(-\frac{i(kp-\omega q)}{\hbar} - \frac{(q-q')^2}{2\sigma^2} - \frac{\sigma^2(p-p')^2}{2\hbar^2}\right)\tilde{\Phi}(k,\omega). \tag{A7}$$

The Gaussian integrations over p and q can now be performed and, after inserting (A3) into the left-hand side of (A2) one finds

$$\int e^{-i(kp'-\omega q')/\hbar}\tilde{A}(k,\omega)\frac{dk\,d\omega}{2\pi\hbar} = \int e^{-i(kp'-\omega q')/\hbar}e^{-\omega^2\sigma^2/2\hbar^2 - k^2/2\sigma^2}\tilde{\Phi}(k,\omega)\frac{dk\,d\omega}{2\pi\hbar}, \tag{A8}$$

and equating the Fourier integrands yields

$$\tilde{\Phi}(k,\omega) = \tilde{A}(k,\omega)\exp\left[\frac{k^2}{2\sigma^2} + \frac{\omega^2\sigma^2}{2\hbar^2}\right], \tag{A9}$$

which defines Φ in terms of the diagonal elements of $\hat{A}$ and the Fourier transform.

The parameter σ in (A9) labels the width of the coherent states used in the Glauber representation, and Φ will, in general, depend on σ. However, because σ is an arbitrary parameter, it cannot enter into any physical results, i.e., if the Glauber representation is used to calculate elements of $\hat{A}$ the result will be independent of σ.

Glauber functions generally depend on both z and z^* in different linear combinations; however, in certain cases the dependence takes a particularly convenient form. Specifically, *if an operator $\hat{A}$ is a function only of $\hat{x}$ (or $\hat{p}$) then the corresponding Glauber function will be a function only of q (or p).*

This proposition can be easily proven through the properties of the Fourier transform or by noting that the Glauber function may be formally expressed as

$$\Phi_A(p,q) = \exp\left[-\tfrac{1}{2}(\partial_p^2 + \partial_q^2)\right]\langle z|\hat{A}|z\rangle \tag{A10}$$

(where the dimensional factors have been suppressed). If $\hat{A}$ can be written as a sum of polynomials in $\hat{p}$ (or $\hat{x}$), then $\langle z|\hat{A}|z\rangle$ will be a sum of polynomials in p (or q) and the only surviving terms from the differential operator will be functions of p (or q). This fact, seemingly obvious but nowhere noted in the literature, hints that the Glauber representation can prove extremely useful even for operators which are highly asymmetric in p and q. For example, the operator $e^{-i\epsilon V(\hat{x})/\hbar}$, commonly obtained from a Trotter product expansion of the evolution operator, can be represented by a Glauber function of one variable, $\Phi_V(q)$.

Because of the above postulate it proves possible to calculate the Glauber function for operators of the form $A(\hat{x})$ [or $B(\hat{p})$] more directly. Assuming (A1) takes the form

$$A(\hat{x}) = \int |z\rangle\Phi(q)\langle z|\frac{d^2z}{\pi}, \tag{A11}$$

we take the diagonal element of both sides in the coordinate basis yielding

$$\begin{aligned} A(x')\delta(0) &= \int |\langle x'|z\rangle|^2\Phi(q)\frac{d^2z}{\pi} \\ &= \int\left\{\frac{e^{-(x'-q)^2/\sigma^2}}{\sqrt{\pi\sigma^2}}\Phi(q)\right\}dq\int\frac{dp}{2\pi\hbar}, \end{aligned} \tag{A12}$$

where the second line follows from the form of the coherent state wavefunction and use of the (p,q) integration measure. Note that in general $A(x') \neq \langle x'|A(\hat{x})|z'\rangle$. Identifying

$$2\pi\hbar\delta(0) = \int dp$$

and Fourier transforming the convolution as in (A2) to (A9) yields

$$\tilde{\Phi}(\omega) = \exp(\omega^2\sigma^2/4\hbar^2)\tilde{A}(\omega), \tag{A13}$$

where $\tilde{A}$ is the Fourier transform of $A(x')$. This expression and (A9) will yield identical Φ's, but (A13) is simpler since it bypasses the calculation of $\langle z'|A(\hat{x})|z'\rangle$.

For the case $A(\hat{x}) = \hat{x}^n$ (A13) can be evaluated analytically to give $\Phi_n(q)$. One finds

$$\begin{aligned} \tilde{A}(\omega) &= \int e^{-i\omega x'/\hbar}x'^n\frac{dx'}{\sqrt{2\pi\hbar}} \\ &= \int\left(i\hbar\frac{\partial}{\partial\omega}\right)^n e^{-i\omega x'/\hbar}\frac{dx'}{\sqrt{2\pi\hbar}} \\ &= \sqrt{2\pi\hbar}\left(i\hbar\frac{\partial}{\partial\omega}\right)^n\delta(\omega). \end{aligned} \tag{A14}$$

Use of relation (A13) then yields

$$\begin{aligned} \Phi_n(q) &= \int e^{i\omega q/\hbar + \omega^2\sigma^2/4\hbar^2}\left(i\hbar\frac{\partial}{\partial\omega}\right)^n\delta(\omega)d\omega \\ &= \left(i\hbar\frac{\partial}{\partial\omega}\right)^n\{e^{i\omega q/\hbar + \omega^2\sigma^2/4\hbar^2}\}_{\omega=0}, \end{aligned} \tag{A15}$$

after n integrations by parts. By rescaling to

$$\lambda = i\omega\sigma/2\hbar,$$

one finds

$$\Phi_n(q) = \left(\frac{\sigma}{2}\right)^n\left(\frac{\partial}{\partial\lambda}\right)^n[e^{2\lambda(q/\sigma)-\lambda^2}]_{\lambda=0}. \tag{A16}$$

The term in braces is quickly recognized as the generating function of the Hermite polynomials; hence,

$$\Phi_n(q) = (\sigma/2)^n H_n(q/\sigma). \tag{A17}$$

A similar expression may be found for $\Phi_n(p)$ in an analogous manner

$$\Phi_n(p) = \left(\frac{\hbar}{2\sigma}\right)^n H_n\left(\frac{\sigma}{\hbar}p\right). \tag{A18}$$

From expressions (A17) and (A18) and the well-known properties of Hermite polynomials, it is apparent that Φ_n will have a leading term like q^n (or p^n) plus other terms proportional to σ^2 [or $(\hbar/\sigma)^2$] to some integer power. But $\sigma^2 \sim \hbar$ so the Glauber function for an operator like $\hat{x}^n$ or $\hat{p}^n$ is given by the corresponding c number plus corrections of higher order in $\hbar$.

[1] R. Feynman, Rev. Mod. Phys. **20**, 367 (1948).
[2] L. S. Schulman, *Techniques and Applications of Path Integration* (Wiley, New York, 1981).
[3] N. Wiener, in *Collected Works* (MIT, Cambridge, MA 1981).
[4] I. M. Gel'fand and A. M. Yaglom, J. Math. Phys. **1**, 48 (1960).
[5] E. J. Heller, Acc. Chem. Res. **14**, 368 (1981).
[6] J. R. Klauder and I. Daubechies, Phys. Rev. Lett. **52**, 1161 (1984); I. Daubechies and J. R. Klauder, J. Math. Phys. **26**, 2239 (1985).
[7] This paper utilizes the convention of normalized, as opposed to holomorphic, coherent states, i.e., $\langle z|z\rangle = 1$ instead of $\langle z|z\rangle = e^{z^*z}$. Although individual formulas will vary depending on conventions, none of the fundamental results of the paper are affected. I wish to thank David G. Boulware for pointing out this possible source of confusion.
[8] The exact origin of this observation is somewhat obscure, but see E. Schrödinger, Ann. Phys. **79**, 489 (1926), and P. A. M. Dirac, Phys. Z. Sowjetunion **3** (1) (1933).
[9] J. R. Klauder, Ann. Phys. **11**, 123 (1960).
[10] J. R. Klauder and B. Skagerstam, *Coherent States* (World Scientific, Singapore, 1985).
[11] S. S. Mizrahi, Physica, **127A**, 241 (1984).
[12] R. Glauber, Phys. Rev. **131**, 2766 (1963).
[13] A. Messiah, *Quantum Mechanics* (Wiley, New York, 1966), Vol. I.
[14] T. L. Marchioro II and T. L. Beck, "The coherent state action in second order form: A generalized Hamilton's principle" (in preparation).
[15] J. R. Klauder, in *Path Integrals*, edited by G. J. Papadopoulos and J. T. Devreese (Plenum, New York, 1989); J. R. Klauder, Phys. Rev. D **19**, 2349 (1979).
[16] M. H. Kalos and P. A. Whitlock, *Monte Carlo Methods* (Wiley, New York, 1986).
[17] J. D. Doll, T. L. Beck, and D. L. Freeman, J. Chem. Phys. **89**, 2170 (1988); J. D. Doll and D. L. Freeman, Adv. Chem. Phys. **73**, 289 (1988); J. D. Doll, D. L. Freeman, and M. J. Gillan, Chem. Phys. Lett., **143**, 277 (1988).
[18] N. Makri and W. H. Miller, Chem. Phys. Lett. **139**, 10 (1987); N. Makri and W. H. Miller, J. Chem. Phys. **89**, 2170 (1988).
[19] C. H. Mak and D. Chandler (preprint).
[20] T. L. Marchioro II and T. L. Beck, "Monte Carlo evaluation of real time coherent state path integrals," submitted to J. Chem. Phys.
[21] Y. Weissman, J. Chem. Phys. **76**, 4067 (1982); Y. Weissman, J. Phys. A: Math. Gen. **16**, 2693 (1983).
[22] D. Huber and E. J. Heller, J. Chem. Phys. **87**, 5302 (1987); D. Huber, E. J. Heller, and R. G. Littlejohn, J. Chem. Phys. **89**, 2003 (1988).

Exactness of the supersymmetric JWKB quantization formula

M. Crescimanno
Department of Physics, 368 LeConte Hall, University of California, Berkeley, California 94720

(Received 9 May 1989; accepted for publication 1 August 1990)

A modification of the Fröman and Fröman [*JWKB Approximation: Contributions to the Theory* (North-Holland, Amsterdam, 1965)] technique is developed in order to study the supersymmetric JWKB formula of Comtet *et al.* [Phys. Lett. B **150**, 159 (1985)]. With this modification, in a direct and nonperturbative manner, it can be proven that this quantization condition is exact for a class of solvable models.

I. INTRODUCTION

Comtet *et al.*[1] have invented a supersymmetry-inspired JWKB quantization formula (hereafter referred to as SJWKB) that, in contrast with the ordinary JWKB quantization condition, is exact for a class of solvable models. For nonsolvable Schroedinger equations the SJWKB condition yields better eigenvalue estimates than the ordinary JWKB condition and has found application in a variety of problems.

In Ref. 1, it was discovered that by writing the Hamiltonian in the so-called "ground state representation;" [2]

$$H = -d_x^2 + \Phi^2(x) - \Phi'(x), \tag{1}$$

and formally treating the Φ' term as $O(\hbar)$ one finds, by a straightforward functional technique, the following quantization condition:

$$\int_a^b \sqrt{E - \Phi^2(x)}\,dx = n\pi\hbar. \tag{2}$$

Many authors have investigated why Eq. (2) is exact for all solvable models.[3–5] In Ref. 3, it was shown that the higher-order corrections (i.e., higher order in $\hbar$) to Eq. (2) vanish for these solvable models.

This work aims at providing a direct and nonperterbative proof of the exactness of the SJWKB condition for these solvable models. We modify the traditional technique that Fröman and Fröman[6] (hereafter referred to as FF) developed for the ordinary JWKB and show conditions for the exactness of the SJWKB quantization condition.[7] This is then used to prove the exactness of the SJWKB formula for several example Hamiltonians. The technique developed here is a novel approach to understanding the SJWKB quantization formula and has certain advantages over other methods, but it seems to go no further in elucidating the connection between exactness, solvability and supersymmetry.[1,8]

In the second section of this paper, we present a brief review of the FF technique for the ordinary JWKB condition. Section III details a modification of the FF technique for supersymmetric Hamiltonians and includes a derivation of the SJWKB quantization condition [Eq. (2)] from a new point of view. In Sec. IV the technique is used to prove the exactness of the SJWKB condition for the harmonic oscillator and the Rosen–Morse potential. Section IV also contains some general remarks about the exactness of the condition for generic potentials at $E = 0$.

II. REVIEW OF THE THEORY OF FRÖMAN AND FRÖMAN

The JWKB quantization condition is exact for the harmonic oscillator and, although many researchers have tried to understand this "coincidence" from various points of view, a proof that was capable of generalization was first formulated in 1964 by Fröman and Fröman.[6] In this section, we will provide a short introduction to the work of FF and demonstrate the exactness of the ordinary JWKB quantization condition for the case of the harmonic oscillator. The reader wishing to consult a more detailed version of this section is entreated to read Refs. 6 and 9 from which most of this has been taken.

The FF technique is a general method for solving equations of the type

$$\frac{d^2\Psi}{dz^2} + Q^2(z)\Psi = 0, \tag{3}$$

where $Q^2(z)$ is a meromorphic function on the complex plane. It is important to note that in our particular application $Q^2(z)$ will be a real-valued function [i.e., $Q^2(z) = -(2m/h^2)(E - V(x))$] on the real axis. This is crucial because the quantization condition is really a result of the hermiticity of the Hamiltonian.

Consider now making the change of variables

$$\Psi = \frac{\varphi(w(z))}{\sqrt{q}}, \quad \text{with } w(z) = \int_{z_0}^{z} q(x)dx, \tag{4}$$

where x and z lie in a connected region of the complex plane and where q is similar to Q except at certain singularities. Since Ψ satisfies Eq. (3), φ solves the following equation:

$$\frac{d^2\varphi}{dw^2} + (1+\epsilon)\varphi = 0, \tag{5}$$

with

$$\epsilon = \frac{Q^2 - q^2}{q^2} + \frac{1}{16q^6}\left(5\left(\frac{dq^2}{dx}\right)^2 - 4q^2\frac{d^2q^2}{dx^2}\right), \tag{6}$$

so, if $\epsilon = 0$ then $\varphi = e^{\pm iw}$ would be exact solutions of Eq. (5). It is easy to see that if ϵ is small for $q^2 = Q^2$ the JWKB wave functions will be good approximations. In order to actually find a q that makes $\epsilon = 0$ one must solve an Airy-like equation. Since it is, in general, not possible to solve Eq. (5) in closed form, consider breaking it into two first-order differential equations by assuming that φ has the following form:

 0022-2488/90/122946-06$03.00

$$\varphi = a_1(w)e^{iw} + a_2(w)e^{-iw} = \mathbf{f}\cdot\mathbf{a},$$

where $\mathbf{f} = (e^{iw}, e^{-iw})$. (7)

It is useful to require that the components of $\mathbf{a}$ satisfy

$$\frac{da_1}{dw}e^{iw} + \frac{da_2}{dw}e^{-iw} = 0. \tag{8}$$

It is difficult to motivate this last step other than by noting that it leads to an immense simplification of the problem. The author suspects however that there may in fact be a geometric way of viewing this constraint. Equation (7) becomes

$$\frac{d\varphi}{dw} = ia_1 e^{iw} - ia_2 e^{-iw}. \tag{9}$$

Combining this with Eq. (5) we find that a_1 and a_2 satisfy the following coupled differential equations:

$$\begin{pmatrix} \frac{da_1}{dw} \\ \frac{da_2}{dw} \end{pmatrix} = \frac{i\epsilon}{2}\begin{bmatrix} 1 & e^{-2iw} \\ -e^{2iw} & -1 \end{bmatrix}\begin{pmatrix} a_1 \\ a_2 \end{pmatrix}. \tag{10}$$

Written vectorially we have $d\mathbf{a}/dw = M(w)\cdot\mathbf{a}$. Note that the matrix $M(w)$ has the special properties $\mathrm{Tr}(M) = 0$, $\mathrm{Det}(M) = 0$ and since $\mathrm{Tr}(M) = 0$, $M(w)$ is an element of a two-dimensional representation of the lie algebra su(2). Define now a matrix F with the property

$$\mathbf{a}(w) = F(w,w_0)\cdot\mathbf{a}(w_0). \tag{11}$$

Then, F solves

$$\frac{dF}{dw} = M\cdot F. \tag{12}$$

Since $F(w_0,w_0) = 1$ and $\mathrm{Tr}(M) = 0$, then $\mathrm{Det}(F) = 1$ is a constant of the flow [F is simply a SU(2) group element]. The solution of Eq. (12) can be written formally in terms of a path-ordered exponential

$$F = P\exp\left(\int_{w_0}^{w} M\,dw\right). \tag{13}$$

It is also obvious that

$$F(z_2,z_0) = F(z_2,z_1)\cdot F(z_1,z_0). \tag{14}$$

Since the w, $\mathbf{a}$, F all depend on q it will be necessary to restrict the domain of these functions to regions in which $\sqrt{q^2}$ is defined. We will therefore take branch cuts on the complex plane where q^2 has either a pole or a zero. For example, for a generic potential (i.e., one in which there is only one well) there will be branch cuts ending on the "classical turning points" labeled in Fig. 1 as a, b. We define q above the cuts by choosing a consistent phase convention. This is done by noting that the phase of q in a region will be related to the angles to the region from the various branch points. For simplicity, consider the case in Fig. 1. Since we may define q up to an overall phase, θ_0, we have

$$q(x) = |q|e^{i\theta_0}e^{i(\theta_a+\theta_b)/2}, \tag{15}$$

where the factor of 1/2 in the exponent is due to the fact that for generic potentials the turning points are simple zeroes of q^2, and we are interested in the phase of $\sqrt{q^2}$. From Fig. 1, we have

region 1: $\theta_a + \theta_b = 2\pi$,

region 2: $\theta_a + \theta_b = \pi$,

region 3: $\theta_a + \theta_b = 0$.

Finally, choosing $\theta_0 = -\pi/2$ will give us what FF call the canonical phase choice for q (in Fig. 1 the χ corresponds only to the phase of q on the real axis).

Since q is real on a segment of the real axis (between the two turning points) then by the Schwarz reflection principle we know

$$q(z^*) = q^*(z). \tag{16}$$

Thus the zero and pole structure of q above the real axis is the mirror reflection of that below the real axis. This symmetry implies that the canonical phases of FF are indeed "canonical;" the *net* contribution to the phase of q along the real axis from the poles and zeros above the real axis exactly cancels that due to the poles and zeros below the real axis. This symmetry is very useful, for it will allow us, in certain cases, to show that only a few poles make a net contribution to certain line integrals. To illustrate this point further consider the q-region associated with the potential $q^2 = E - U_0\,\mathrm{sech}^2(z)$. A choice of branch cuts is shown in Fig. 2. The poles of q^2 are at $z = i\pi(n + \frac{1}{2})$ and the zeros are $z = a,b + i\pi n$. Note that for the path Γ in the figure the integral of a function depending on q can receive a net contribution only from the singularity at $i\pi/2$.

Now let us comment on Eq. (12). Thinking of $M(w)$ as a two-dimensional vector potential (with components depending on w and $\bar{w}$ in general), we may then understand F as the Wilson line of $M(w)$. Since $M(w)$ is analytic on the region in which q is defined, the non-Abelian Stokes theorem assures us that the value of the Wilson line, F, is a functional

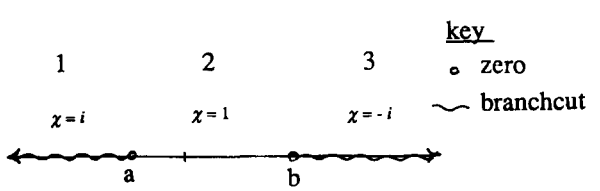

FIG. 1. Choice of canonical phases of $q = \chi|q|$ on real axis for regions 1, 2, and 3.

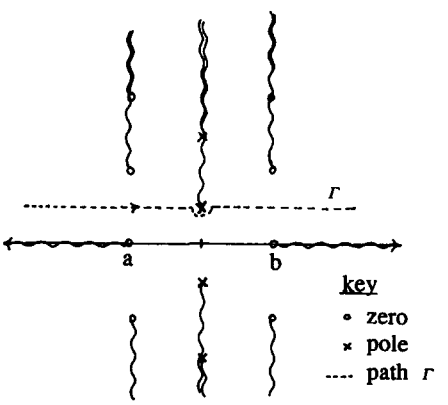

FIG. 2. The q region of potential $q^2 = E - U_0\,\mathrm{sech}^2(z)$.

of $M(w)$ which depends only on the homotopy class of the path of integration.

In general, it is both the hermiticity of the Hamiltonian and the L^2 norm properties of the solutions that are responsible for the existence of quantization conditions. We now discuss the consequences of hermiticity of the Hamiltonian in this formulation. For the following, we restrict our attention to the real axis. A simple consequence of hermiticity is that for every solution Ψ, the complex conjugate, Ψ^*, is also a solution. Since the wave function is

$$\Psi = \frac{a_1(w)e^{iw} + a_2(w)e^{-iw}}{\sqrt{q}} = \mathbf{f}'\cdot\mathbf{a}, \tag{17}$$

where we for convenience define,

$$\mathbf{f}' = \mathbf{f}/\sqrt{q}, \tag{18}$$

then $\Psi^* = \mathbf{f}'^*\cdot\mathbf{a}^*$. $\mathbf{f}'$ is so simple that, with a slight abuse of notation, we may define a matrix, B, that applied to $\mathbf{f}'$ yields $\mathbf{f}'^*$,

$$\mathbf{f}'^* = \mathbf{f}'\cdot B, \quad \mathbf{f}'^{**} = \mathbf{f}' \text{ so } B^*B = 1,$$
$$\mathrm{Det}(B) = \pm 1. \tag{19}$$

Note that * denotes complex conjugation and not the Hermitian conjugate. Furthermore, since we are studying Ψ on the real axis, we may require that,

$$\frac{d\mathbf{f}'^*(x)}{dx} = \frac{d\mathbf{f}'(x)}{dx}\cdot B(x). \tag{20}$$

For example, we may write B in such a way that it depends only on which region x is in but not on its actual value. We may do this because the w integral will receive contributions of fixed phase in each of the regions 1,2,3. Explicitly, these matrices are

$$B(x) = \frac{q(x)}{|q(x)|}\begin{bmatrix} 0 & e^{2\,\mathrm{Im}(w)} \\ e^{-2\,\mathrm{Im}(w)} & 0 \end{bmatrix}, \tag{21a}$$

for regions in which $q^2(x) > 0$ and

$$B(x) = \frac{q(x)}{|q(x)|}\begin{bmatrix} e^{-2i\,\mathrm{Re}(w)} & 0 \\ 0 & e^{2i\,\mathrm{Re}(w)} \end{bmatrix}, \tag{21b}$$

for regions in which $q^2(x) < 0$. The **a** vector in our wave function, $\Psi = \mathbf{f}'\cdot\mathbf{a}$, was chosen so that,

$$\frac{d\Psi}{dx} = \frac{d\mathbf{f}'}{dx}\cdot\mathbf{a}. \tag{22}$$

Since $\Psi^* = \mathbf{f}'\cdot B\cdot\mathbf{a}$ and $dB/dx = 0$, then,

$$\frac{d\Psi^*}{dx} = \frac{d\mathbf{f}'}{dx}\cdot B\cdot\mathbf{a}^* \tag{23}$$

so $B\cdot\mathbf{a}^*$ behaves like an "**a**" vector for Ψ^*. It then follows, using Eq. (11), that

$$B(x_1)\cdot\mathbf{a}^*(x_1) = F(x_1,x_2)\cdot B(x_2)\cdot\mathbf{a}^*(x_2). \tag{24}$$

Now, using the fact that $B^* = B^{-1}$ and again making use of relation Eq. (11), we have the identity,

$$F(x_1,x_2) = B(x_1)\cdot F^*(x_1,x_2)\cdot B^*(x_2). \tag{25}$$

This conjugation symmetry relationship leads to connections between the entries of the F matrix in different regions. One formula of particular interest is that in which both x_1 and x_2 lie in regions in which $q^2(x) < 0$. Suppose furthermore that the x_1 and x_2 lie in different regions, i.e., straddling the classical allowed region. Then, writing out Eq. (25) for the component F_{22}, we find

$$F_{22}(x_1,x_2) = \frac{q(x_1)/|q(x_1)|}{q(x_2)/|q(x_2)|}e^{-2i\,\mathrm{Re}(w(x_1)-w(x_2))} \times F^*_{22}(x_1,x_2). \tag{26}$$

This is the quantization condition. Since for a single well-generic potential we have chosen the x_1 and x_2 to lie in the separate classically forbidden regions, our phase convention yields

$$\frac{q(x_1)/|q(x_1)|}{q(x_2)/|q(x_2)|} = -1. \tag{27}$$

Furthermore, since the w receives only imaginary contributions from the integration in the regions 1 and 3,

$$\mathrm{Re}(w(x_1) - w(x_2)) = \int_a^b |q|dx. \tag{28}$$

Finally, using this in Eq. (26) yields,

$$F_{22}(x_1,x_2) = \exp\left[-2i\int_a^b |q|dx + i\pi\right]F^*_{22}(x_1,x_2), \tag{29}$$

and so,

$$F_{22}(x_1,x_2)\exp\left[i\int_a^b |q|dx - \frac{i\pi}{2}\right] \tag{30}$$

is real. This implies that,

$$\int_a^b |q|dx = \left(n+\frac{1}{2}\right)\pi - \arg(F_{22}(x_1,x_2)), \tag{31}$$

so we readily see that the ordinary JWKB quantization condition will be exact if F_{22} is real.

In Eq. (31), x_1 and x_2 are arbitrary and so we may conveniently let x_1 and x_2 approach $-\infty$ and $+\infty$, respectively. In regions 1 and 3, the e^{iw} component is monotonically increasing (since we are considering only generic potentials) and in order to have a normalizable JWKB wave function a_1 must vanish sufficiently quickly as x approaches $\pm\infty$. Since $a_2(\infty) \neq 0$ (else Ψ would vanish identically), we can use $a_2(+\infty)$ to give $a_1(z)$ and $a_2(z)$ by using Eq. (11). It then follows that

$$\frac{a_1(z)}{a_2(z)} = \frac{F_{12}(z,+\infty)}{F_{22}(z,+\infty)}. \tag{32}$$

Similarly, we can start with $a_2(-\infty)$ and find $a_1(z)$ and $a_2(z)$, thus

$$\frac{F_{12}(z,-\infty)}{F_{22}(z,-\infty)} = \frac{F_{12}(z,+\infty)}{F_{22}(z,+\infty)}. \tag{33}$$

From the multiplication law of the F matrices [the group law Eq. (14)], it is obvious that $F_{12}(-\infty,+\infty) = 0$. Combining this and the above formulas yields an asymptotic statement on F_{22}:

$$F_{22}(-\infty,+\infty) = \frac{F_{22}(z_0,+\infty)}{F_{22}(z_0,+\infty)}. \tag{34}$$

FF have developed expansions for the individual components of the F matrices. For our purposes here, we will not

need to make use of these expansions but refer the interested reader to Ref. 6 and also to Ref. 9 for several nontrivial examples. Instead, we would like to conclude this section with a demonstration of this technique for the harmonic oscillator.

For the harmonic oscillator let $Q^2 = q^2 = E - Az^2$. Then, the complex q plane looks like Fig. 1 and has only two branch cuts on the entire plane originating at $\pm\sqrt{A/E}$. Since $M(z)$ is analytic in the entire upper half-plane, we may deform a contour running along the real axis from $-\infty$ to $+\infty$ into a semicircle, S, of some large radius, R. Note now that since along S, $\epsilon \approx \text{const.}/R^4$ then in the limit of large R the contribution to the path-ordered integration along S vanishes. Then, by Eq. (34) with initial condition $F(z_0,z_0) = 1$, we have $F_{22}(-\infty,+\infty) = 1$ and so the JWKB quantization condition is exact for the harmonic oscillator.

III. SUPERSYMMETRIC JWKB FORMULA

This section contains the main result of this paper; a modification of the FF technique particular to supersymmetric Hamiltonians.[7] We will try to follow the notation of Sec. II as closely as possible. To begin with, consider the convention for the Hamiltonian in Eq. (1). Once again, we make the ansatz

$$\Psi = \varphi/\sqrt{q}, \tag{35}$$

where now the $q^2 = E - \Phi^2$. As in Sec. II, we write φ as a quadratic form of the vectors $\mathbf{f}$ and $\mathbf{a}$,

$$\varphi = \mathbf{f}\cdot G\cdot\mathbf{a}, \tag{36}$$

where the matrix G is

$$G = \begin{bmatrix} e^{i\sigma} & 0 \\ 0 & e^{-i\sigma} \end{bmatrix} \quad \text{with} \quad \sigma = \int_{z_0}^{z} \frac{\Phi'\,dz}{2q}. \tag{37}$$

From Eq. (1) and Eq. (35), we find that φ solves

$$\frac{d^2\varphi}{dw^2} + \left(1 + \epsilon + \frac{\Phi'}{q^2}\right)\varphi = 0 \tag{38}$$

with

$$\epsilon = \frac{1}{16q^6}\left(5\left(\frac{dq^2}{dx}\right)^2 - 4q^2\frac{d^2q^2}{dx^2}\right).$$

Suppose we now require the vector $G\cdot\mathbf{a}$ to have properties similar to the $\mathbf{a}$ vector of Sec. II. Then,

$$\frac{d\varphi}{dw} = \frac{d\mathbf{f}}{dw}\cdot G\cdot\mathbf{a} \quad \text{and} \quad \mathbf{f}\cdot\left(\frac{dG}{dw}\cdot\mathbf{a} + G\cdot\frac{d\mathbf{a}}{dw}\right) = 0, \tag{39}$$

and so,

$$\frac{d^2\varphi}{dw^2} = \frac{d^2\mathbf{f}}{dw^2}\cdot G\cdot\mathbf{a} + \frac{d\mathbf{f}}{dw}\cdot\left(\frac{dG}{dw}\cdot\mathbf{a} + G\cdot\frac{d\mathbf{a}}{dw}\right). \tag{40}$$

Since $d^2\mathbf{f}/dw^2 = -\mathbf{f}$, we can rewrite Eq. (38) as

$$\frac{d\mathbf{f}}{dw}\cdot\left(\frac{dG}{dw}\cdot\mathbf{a} + G\cdot\frac{d\mathbf{a}}{dw}\right) + \left(\epsilon + \frac{\Phi'}{q^2}\right)\mathbf{f}\cdot G\cdot\mathbf{a} = 0. \tag{41}$$

Finally, using the fact that $d\mathbf{f}/dw\cdot dG/dw\cdot\mathbf{a} = -(\Phi/q^2)\mathbf{f}\cdot G\cdot\mathbf{a}$ we can find, as was done in the section on the FF technique for the ordinary JWKB, the matrix that defines the evolution of the $\mathbf{a}$ vector. We have

$$\frac{d\mathbf{a}}{dw} = M_s\cdot\mathbf{a}, \tag{42}$$

with

$$M_s = \frac{i\epsilon}{2}\begin{bmatrix} 1 & ke^{-2i(w+\sigma)} \\ -ke^{2i(w+\sigma)} & -1 \end{bmatrix}$$

$$\text{and} \quad k = 1 + \frac{\Phi'}{q^2\epsilon}. \tag{43}$$

The matrix M_s is more complicated than M of Sec. II. Note once again that $\mathrm{Tr}(M_s) = 0$ and so M_s is again an element of su(2). One striking difference between M and M_s is that $\mathrm{Det}(M_s) \neq 0$ almost everywhere. We can once again define an F matrix that defines the evolution of the $\mathbf{a}$ vector as a linear map,

$$\mathbf{a}(w) = F(w,w_0)\cdot\mathbf{a}(w_0). \tag{44}$$

Then, of course,

$$\frac{dF}{dw} = M_s\cdot F. \tag{45}$$

Again, since the $\mathrm{Tr}(M_s) = 0$ and F is solved with the initial condition $F(w_0,w_0) = 1$, we have $\mathrm{Det}(F) = 1$.

We now explore the consequences of the hermiticity of the Hamiltonian. Let us start as we did in Sec. II by again focusing our attention on the real axis. We can define a matrix B that is the linear operator of complex conjugation of the $\mathbf{f}$ vector, exactly as before,

$$\mathbf{f}'^* = \mathbf{f}'\cdot B. \tag{46}$$

The matrix B is of course the same as in Sec. II. Since our complete supersymmetric wave function is $\Psi = \mathbf{f}'\cdot G\cdot\mathbf{a}$, we have that $\Psi^* = \mathbf{f}'\cdot B\cdot G^*\cdot\mathbf{a}^*$ and so,

$$\frac{d\Psi^*}{dx} = \frac{d\mathbf{f}'}{dx}\cdot B\cdot G^*\cdot\mathbf{a}^*. \tag{47}$$

The factor $B\cdot G^*\cdot\mathbf{a}^*$ behaves like the "$G\cdot\mathbf{a}$" vector of Ψ^*. Let us denote this "$G\cdot\mathbf{a}$" vector of Ψ^* by $G\cdot\mathbf{b}$. We then have

$$G\cdot\mathbf{b} = B\cdot G^*\cdot\mathbf{a}^*$$

and so,

$$\mathbf{b} = CG^{-1}\cdot B\cdot G^*\cdot\mathbf{a}^*. \tag{48}$$

Using the above results in $\mathbf{b}(x_1) = F(x_1,x_2)\cdot\mathbf{b}(x_2)$ yields

$$F(x_1,x_2) = G^{-1}\cdot B\cdot G^*(x_1)\cdot F^*(x_1,x_2)\cdot G^{*-1}\cdot B\cdot G(x_2). \tag{49}$$

We can now derive the quantization condition by once again requiring x_1 and x_2 to be in seperate classically forbidden regions, i.e., those in which $q^2(x) < 0$. Being careful about taking complex conjugates of the elements in the matrix G (it is important to note that its phases vary in the different regions) and concentrating on just the F_{22} element of the F matrix, we find

$$F_{22}(x_1,x_2) = \frac{q(x_1)/|q(x_1)|}{q(x_2)/|q(x_2)|}RF^*_{22}(x_1,x_2), \tag{50}$$

with

$$R = \exp(i[(\sigma_1 + \sigma_1^*) - (\sigma_2 + \sigma_2^*) + 2\,\mathrm{Re}(w(x_1) - w(x_2))]),$$

where σ_1 and σ_2 are $\sigma(x_1)$ and $\sigma(x_2)$, respectively. With our chosen x_1 and x_2 the prefactor is,

$$\frac{q(x_1)/|q(x_1)|}{q(x_2)/|q(x_2)|} = -1. \tag{51}$$

Although it looks as though the factor involving σ might depend on the points x_1 and x_2, we note that $\sigma + \sigma^*$ is purely real and so beyond the classical turning points, it receives no contribution. This means that the exponent involving the σ's is

$$(\sigma_1 + \sigma_1^*) - (\sigma_2 + \sigma_2^*) = \int_a^b \left(\frac{\Phi'}{q}\right) dx = \pi, \tag{52}$$

and, hence, $F_{22}(x_1,x_2)\exp[i\int_a^b |q|dx]$ is real, which implies that,

$$\int_a^b |q|dx = n\pi - \arg(F_{22}(x_1,x_2)). \tag{53}$$

This is a generalization of the quantization condition [Eq. (2)] of Ref. 1. Equation (53) shows that the SJWKB condition [Eq. (1)] will be exact if $F_{22}(x_1,x_2)$ is real. Again, we may use the arguments of the previous section regarding the normalizability of the wave functions to recast this result into the form

$$\int_a^b |q|dx = n\pi - \arg(F_{22}(-\infty,+\infty)), \tag{54}$$

where $F_{22}(-\infty,+\infty)$ is given by

$$F_{22}(-\infty,+\infty) = F_{22}(z_0,+\infty)/F_{22}(z_0,+\infty). \tag{55}$$

As mentioned earlier, FF have developed a series of bounds for the elements of the F matrix. In those formulas, heavy use was made of the specific form of the M matrix in Eq. (10). For our more complicated matrix M_s [see Eq. (43)], the same method of constructing bounds seems overly complicated and will not be necessary for this paper. Instead, simple analytical methods will be used to determine the phase of $F_{22}(-\infty,+\infty)$ of Eq. (54) in the applications in the next section.

IV. SEVERAL EXAMPLES

In this section, we will demonstrate the exactness of the SJWKB quantization condition for several solvable systems by proving that $F_{22}(-\infty,+\infty)$ is real. The examples particularly amenable to exposition are the harmonic oscillator and the Rosen–Morse potentials. Included in this section are also some remarks about the consistency of this method for the $E = 0$ case.

A. The supersymmetric harmonic oscillator

Consider the case $\Phi = \omega x$. The potential is then that of a harmonic oscillator. For this Φ, the q plane is shown in Fig. 3. Note that, as in the case of the analysis of the ordinary JWKB for the harmonic oscillator, q has no other branch cuts on the complex plane other than those associated with the classical turning points. Again, consider integrating Eq. (45) along the contour S; a semicircle of radius R lying entirely in the upper half-plane. Now, using Eq. (38) for ϵ and studying the asymptotics as $R \to \infty$ of ϵ and ϵk, we again see

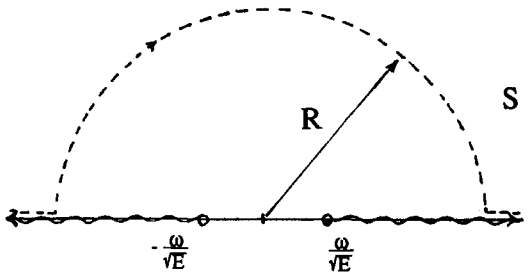

FIG. 3. The q plane for supersymmetric harmonic oscillator.

that (subject to the initial conditions $F(z_0,z_0) = 1$) that $F_{22}(-\infty,+\infty)$ is 1, thereby demonstrating that the SJWKB quantization condition is exact for the Harmonic oscillator.

B. The supersymmetric Rosen–Morse potential

The Rosen–Morse Hamiltonian is

$$H = -d_x^2 - U_0 \operatorname{sech}^2(\alpha z). \tag{56}$$

In this case, $\Phi = (A/2)\tanh(\alpha z)$ with $A = \sqrt{\alpha^2 + 4U_0} - \alpha$. Therefore, $\Phi' = (\alpha A/2)\operatorname{sech}^2(\alpha z)$ and so Eq. (38) for ϵ reads

$$\epsilon = [\alpha^2 s/A^2(r+s)^3][s^2 + (6r+1)s - 4r], \tag{57}$$

where $s = \operatorname{sech}^2(\alpha z)$ and $r = 4E/A^2$. Note that as $z \to \pm\infty$ along the real axis $s \to 0$ and $\epsilon \to 0$ exponentially, so the parameter integrals of ϵq and $\epsilon k q$ will not diverge at $\pm\infty$. The region of analyticity of q is shown in Fig. 4. Again, by the Schwarz reflection principle, we are assured that we may choose the phases of q along the real axis to be the canonical ones discussed earlier. Also note that along the dotted contour Γ' in Fig. 4, only the pole at $i\pi/2\alpha$ makes a net contribution to the choice of phase of q along Γ'. Since the pole at $i\pi/2\alpha$ is a double pole of q^2, the phase of q along the right-hand part of Γ' is

$$q = |q|e^{i\theta_0}e^{i(0)(-2/2)} = |q|e^{i\theta_0}, \tag{58}$$

and that along the left-hand part is

$$q = |q|e^{i\theta_0}e^{i(-\pi)(-2/2)} = -|q|e^{i\theta_0}. \tag{59}$$

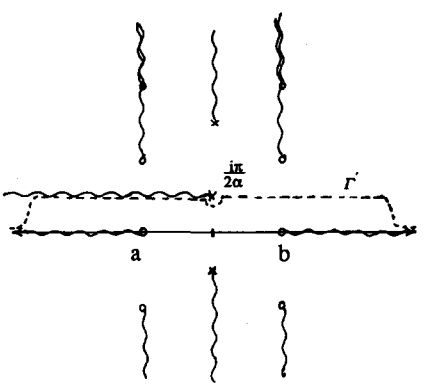

FIG. 4. The q plane for supersymmetric Rosen–Morse potential.

Now, using the canonical phase convention for θ_0, the phase of q along Γ' is

$$q = i|q| \quad \text{on region 1,}$$
$$q = -i|q| \quad \text{on region 2.} \tag{60}$$

This implies that w and σ are pure imaginary functions along Γ'. Then, $e^{i(w+\sigma)}$ is a real function of x, the coordinate along the path Γ'. It is also rather easy to check that both ϵ and ϵk are real bounded functions of w along Γ'. Thus the path-ordered integral of the matrix M_s from $-\infty$ to $+\infty$ can be carried out along Γ' except for little "jogs" to the real axis at $\pm\infty$. Now, since both ϵ and ϵk vanish at $\pm\infty$ and since w and σ are also well-behaved asymptotically, the "jogs" at $\pm\infty$ make no contribution to $F(-\infty, +\infty)$. Then, using the fact that qM_s is real, we have that

$$\frac{d(F-F^*)}{dx} = qM_s(x)\cdot(F-F^*), \tag{61}$$

and since, choosing some basepoint x_0 along Γ' and the initial condition $F(x_0,x_0) = 1$ implies $(F-F^*)(x_0,x_0) = 0$, the imaginary part of the F matrix remains zero along Γ' (the singularity in q^2 at $i\pi/2\alpha$ is benign in qM_s, as may be checked by the reader). Therefore, by Eq. (55), $F_{22}(-\infty, +\infty)$ is real and so the SJWKB quantization condition is exact.

In these two cases, the exactness of the SJWKB has been so easy to understand with this technique that the proof was carried out with virtually no calculation. Note that for the nonsolvable potential $az^2 + \lambda z^4$ the branch cuts that are off the real axis in the q plane obviate the use of arguments similar to those made for the harmonic oscillator. Although we have not rigorously shown that $F_{22}(-\infty, +\infty)$ is not real in this case, it is no surprise that the SJWKB is not exact for this potential (a fact that the interested reader may verify in perturbation theory in λ). In general, we would expect the $F_{22}(-\infty, +\infty)$ to depend on E. It might be possible to find asymtotic forms for $F_{22}(-\infty, +\infty)$ as a function of E, but we leave that possibility to future development of this technique.

We now wish to make a few simple comments about the SJWKB condition for a generic potential for $E = 0$. For Eq. (2) with $n = 0$, $E = 0$ is a consistent solution because for a generic potential as $E \to 0$ the distance between the classical turning points vanishes. In this case, the q plane looks like that of Fig. 5. Note that now $q^2 = 0$ implies that $\Phi = 0$ so the zero is at least a double zero. If the zero is a double zero, the phases in regions 1 and 3 are again i and $-i$, respectively. This means that the matrix qM_s is real, and if the integrals of ϵ and ϵk are defined, then $F_{22}(-\infty, +\infty)$ is, by arguments similar to those used above, well defined and real. Thus the SJWKB is again exact at for any generic potential at $E = 0$. Let us emphasize that this "exactness at $E = 0$" for generic potentials is not deep but is simply a check of the consistency of this technique.

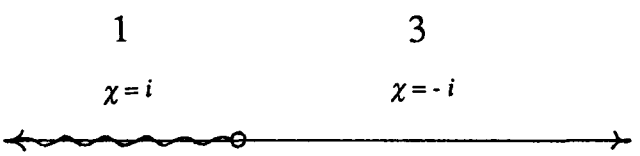

FIG. 5. The q plane for generic potential at $E = 0$.

V. CONCLUSION

We have generalized the techniques of FF to the case of supersymmetric Hamiltonians. This method yields a derivation of a generalization, Eq. (54), of the SJWKB quantization condition of Comtet *et al.*[1] It furthermore makes the exactness of a class of solvable models easy to understand. It is a fully nonpeturbative technique. The deeper connections between solvability, exactness and supersymmetry remain to be explored.

ACKNOWLEDGMENTS

The author would like to thank the physics department at Princeton for hospitality during spring 1985 when this work was completed, especially J. Harvey for having brought this topic to my attention and E. Lieb for his insight and interest in this problem. The author also gladly acknowledges a useful conversation with D. K. Campbell and M. M. Nieto. The author also thanks K. Bardakci and M. B. Halpern for their interest in this problem and S. Hsu and A. Landsberg for helpful comments.

The author's primary source of financial support is an N.S.F. Graduate Fellowship. This work is supported in part by N.S.F. Grant no PHY 85-15857.

[1] A. Comtet, A. Bandrauk, and D. Campbell, Phys. Lett. B **150**, 159 (1985).
[2] E. Gozzi, Phys. Lett. B **129**, 432 (1983).
[3] K. Raghunathan, M. Seetharaman, and S. S. Vasan, Phys. Lett. B **188**, 351 (1987); see also, S. S. Vasan, M. Seetharaman, and K. Raghunathan, J. Phys. A **21**, 1897 (1988).
[4] R. Dutt, A. Khare, and U. P. Sukatme, Phys. Lett. B **181**, 295 (1986).
[5] S. H. Fricke, A. B. Balantekin, P. J. Hatchell, and T. Uzer, Phys. Rev. A **37**, 2797 (1988).
[6] N. Fröman and P. O. Fröman, *JWKB Approximation; Contributions to the Theory* (North-Holland, Amsterdam, 1965).
[7] M. Crescimanno, Ph.D, Thesis, Princeton University, Princeton, NJ, 1985 (unpublished).
[8] M. M. Nieto, Phys. Lett. B **145**, 208 (1984).
[9] C. Rosenzweig and J. Krieger, J. Math. Phys. **9**, 849 (1968).

Semiclassical structure of trace formulas

Robert G. Littlejohn[a)]
Theoretical Physics Institute, School of Physics and Astronomy, 116 Church Street S.E., Minneapolis, Minnesota 55455

(Received 14 February 1990; accepted for publication 18 July 1990)

Trace formulas provide the only general relations known connecting quantum mechanics with classical mechanics in the case that the classical motion is chaotic. In particular, they connect quantal objects such as the density of states with classical periodic orbits. In this paper, several trace formulas, including those of Gutzwiller, Balian and Bloch, Tabor, and Berry, are examined from a geometrical standpoint. New forms of the amplitude determinant in asymptotic theory are developed as tools for this examination. The meaning of caustics in these formulas is revealed in terms of intersections of Lagrangian manifolds in phase space. The periodic orbits themselves appear as caustics of an unstable kind, lying on the intersection of two Lagrangian manifolds in the appropriate phase space. New insight is obtained into the Weyl correspondence and the Wigner function, especially their caustic structures.

I. INTRODUCTION

This paper concerns the trace formulas of Gutzwiller[1] and Balian and Bloch,[2] which expresses the density of states of a bound quantal system as a sum over the periodic orbits of the corresponding classical system, and closely related trace formulas, such as that of Tabor[3] for the density of quasistates in a time-periodic system, and that of Berry[4] for the scars of Wigner functions in phase space. The purpose of this paper is to explore the semiclassical structures of such formulas, i.e., the geometrical objects in the classical phase space associated with them and the interplay between these objects and the corresponding wave fields. The basic framework for such a study is the theory of Maslov,[5] which associates the asymptotic properties of wave fields with Lagrangian manifolds in phase space. The application of this theory to trace formulas is unusual and interesting, and involves several novel elements.

The trace formulas considered here are significant because they are among the very few theoretical results of any generality that connect quantum mechanics with classical mechanics in the case that the classical motion is chaotic. These formulas work equally well for classically integrable systems, and in that case they have been shown by Berry and Tabor[6] to be equivalent to the usual Bohr–Sommerfeld methods (appropriately called "torus quantization" by Berry).[7] However, the usual methods fail for chaotic systems, mainly because of the lack of well behaved, invariant Lagrangian manifolds in phase space. For this reason, it is easy to get the impression that standard wave asymptotics or WKB theory does not apply at all to chaotic systems. We will show, however, that there is a rich geometrical structure associated with trace formulas, even in chaotic systems. For example, it turns out that Gutzwiller's periodic orbits define a certain, unstable kind of caustic, represented by the intersection of two Lagrangian manifolds in the appropriate phase space. We will also show that the density of states can itself be regarded as a wave function, in which the periodic orbit terms correspond to the branches of a WKB formula as in standard asymptotic theory.

We will be especially interested in Gutzwiller's[1] method of deriving his trace formula, since it connects most immediately with the geometry of phase space. The methods of Tabor[3] and Berry[4] for deriving their trace formulas are essentially identical to Gutzwiller's, and may be described as a determined application of the stationary phase approximation. Reviews of the trace formula, such as those of Berry[7] and Ozorio de Almeida,[8] have basically followed these methods. Balian and Bloch[2] also used similar methods, but built them around an elegant formalism based on Laplace transforms in the quantity $1/\hbar$. From the standpoint of phase space geometry, however, the work of Balian and Bloch does not seem as useful a place to start as that of Gutzwiller. In addition, Balian and Bloch make some assumptions about classical mechanics which are certainly not correct, such as the idea that one can avoid caustics and multivalued solutions to the Hamilton–Jacobi equation simply by invoking complex energies. We will not be concerned at all in this paper with the Selberg trace formula,[9] although it continues to attract a great deal of interest, because it is an exact result obtained by special methods for a special system, and we are interested here in asymptotic results of general applicability.

One of the original goals of this work was to simplify Gutzwiller's derivation of his trace formula, especially the difficult manipulations of amplitude determinants. This led to a general examination of amplitude determinants and WKB wave functions, which are discussed in Sec. II. We are careful to distinguish a particular solution of the Hamilton–Jacobi equation from a complete solution, the former being represented geometrically by a single, isolated Lagrangian manifold in phase space, and the latter by a foliation of phase space into Lagrangian manifolds. In the case of a particular solution, the amplitude determinant represents a density on the isolated Lagrangian manifold, according to the standard picture of Maslov;[5] but in the case of a complete solution, the provocative work of Miller[10] reveals connections between

[a)] Permanent address: Department of Physics, University of California, Berkeley, CA 94720.

 0022-2488/90/122952-26$03.00

the solution of the Hamilton–Jacobi equation and that of the amplitude transport equation, based on concepts of unitarity. This is the idea we pursue in Sec. II, in order to develop a representation of amplitude determinants in terms of complete sets of classically commuting observables. An outgrowth of this study is what appears to be a new expression for amplitude determinants, in terms of Poisson brackets connecting two complete sets of commuting observables. This expression is given below in Eq. (2.16), and part of its value to us is that it casts amplitude determinants into a form whose invariant meaning in phase space is manifest.

Section III consists of a study of the role of propagators and Green's functions in Gutzwiller's derivation of his trace formula, based on standard asymptotic theory and on some of the notions developed in Sec. II. For example, we identify some of the manifolds, densities, and complete sets of commuting observables used in Gutzwiller's manipulations of propagators and Green's functions. We also show that there is a close relationship between the energy-dependent Green's function and an object best described as a propagator for evolution in the energy shell.

The main theme of Sec. IV is the use of the Hilbert–Schmidt scalar product of operators to interpret traces in terms of the asymptotic theory normally used for wave functions. This approach leads naturally to a doubled phase space with a doubled version of the Poisson bracket. The doubled phase space is a well-known device, explained, for example, by Abraham and Marsden,[11] for representing canonical transformations, and in this role it has an interesting interplay with the asymptotics of matrix elements of operators, especially unitary operators. The accumulation of evidence presented in Sec. IV is, I believe, convincing that the doubled phase space is the necessary and correct medium for understanding the geometrical aspects of trace formulas, as well as many features of the Wigner function and Weyl correspondence.

Finally, in Sec. V, some conclusions are drawn and suggestions made for future work.

The work reported on in this paper began as an attempt to simplify Gutzwiller's derivation of his trace formula, and to reveal whatever deeper structures might underlie it. In the process of carrying out this program, however, there emerged much of the geometrical structure belonging to the mathematical theory of Fourier integral operators. For example, the idea of using the doubled phase space with the doubled symplectic form of Eq. (4.10) is one of the founding ideas of this theory, and is due to Hörmander.[12] Later, an apparently independent derivation of Gutzwiller's trace formula was made by Duistermaat and Guillemin,[13] who interpreted the periodic orbits in terms of the intersections of a given Lagrangian manifold with the diagonal, representing the identity. Also, the role of intersecting Lagrangian manifolds in geometric quantization is discussed by Blattner and Kostant.[14] These and other issues are developed and discussed at greater length by Weinstein[15] and Guillemin and Sternberg.[16]

Unfortunately, these mathematical developments have taken place in almost complete isolation from the more applied community, which is interested in trace formulas as a tool for understanding wave chaos. Therefore, this paper should not be taken as the development of new mathematics (although it is hoped that experts in the relevant fields will find some useful ideas), but rather as an application of existing mathematics to some problems in wave chaos.

II. STRUCTURE OF SEMICLASSICAL WAVE FUNCTIONS AND MATRIX ELEMENTS

In this section we investigate the phase space structure associated with semiclassical wave functions and matrix elements. We are especially interested in the general solution of the Hamilton–Jacobi equation, expressed in terms of complete sets of commuting observables, and in the restrictions which a given solution of the Hamilton–Jacobi equation imposes on the solutions of the amplitude transport equation. This line of inquiry leads naturally to Miller's formula for semiclassical wave functions, Eq. (2.8) below. Although many wave functions of interest in semiclassical mechanics fall under Miller's formula, not all of them do, and we provide the appropriate generalization, given in Eq. (2.12).

Next we turn to a reformulation of Miller's formula, in which the amplitude determinant is expressed purely in terms of Poisson brackets, thereby revealing its invariance under canonical transformations in a manifest manner. The new formulation of Miller's formula is displayed below in Eq. (2.16). The Poisson bracket version of the amplitude provides interesting perspectives on caustics, which we discuss. The Miller formula must be modified in the presence of caustics, with most of the changes occurring in the amplitude determinant. There are as many modifications as there are caustic types; these include the standard catastrophes,[17] but also include caustics that are not catastrophes because they are not stable with respect to small perturbations. The latter class of caustics is especially of interest to us, because the periodic orbits in the Gutzwiller trace formula are precisely caustics of this kind. We work out the appropriate modifications to the Miller formula in the presence of such caustics, again expressing the amplitude in terms of Poisson brackets, and display the result in Eq. (2.27).

A. Semiclassical wave functions

Let us take the ordinary, time-independent Schrödinger equation $\hat{H}\psi = E\psi$ as a specific example to work with, although later we may want to replace $\hat{H}$ by some other operator. We apply standard, multidimensional WKB theory to this equation, and write $\psi(x) = \Omega(x)e^{iS(x)/\hbar}$, giving us the Hamilton–Jacobi equation for the action $S(x)$,

$$H\left(x, \frac{\partial S}{\partial x}\right) = E, \tag{2.1}$$

and the amplitude transport equation for the amplitude $\Omega(x)$,

$$\frac{\partial}{\partial x}\left[\Omega(x)^2 \frac{\partial H(x,p(x))}{\partial p}\right] = 0, \tag{2.2}$$

where $p(x) = \partial S(x)/\partial x$. In these equations we write simply x for the configuration space coordinates $(x_1, \ldots, x_f)$, where f is the number of degrees of freedom, making no attempt to distinguish notationally between the one-dimensional and

multidimensional cases. The required contractions and scalar products will usually be obvious; where they are not, we will explicitly insert indices. We will use a similar notation for the momentum p and velocity v and other variables that are naturally interpreted as f-vectors. We will also sum over repeated indices, except as noted.

A particular solution $S(x)$ of the Hamilton–Jacobi equation is always the generating function of an invariant Lagrangian manifold in phase space. Lagrangian manifolds are f-dimensional surfaces in the $2f$-dimensional phase space on which the symplectic form vanishes; their theory is discussed by Arnold,[18] and their use in WKB theory is explained by Maslov and Fedoriuk,[5] Percival,[19] and Delos.[20] Here, we will simply note a few basic facts about them and about their generating functions. First, a Lagrangian manifold is invariant, i.e., mapped into itself by the flow, if and only if it is a subset of some energy shell $H(x,p) = E$. A convenient way of constructing an invariant Lagrangian manifold is to choose an arbitrary, $(f-1)$-dimensional Lagrangian manifold L_0 in a surface of section, regarded as a phase space of $(f-1)$ degrees of freedom in its own right, and to let L_0 move with the flow into the energy shell, sweeping out an f-dimensional manifold in the energy shell. It turns out that this manifold is Lagrangian in the full $2f$-dimensional phase space. The generating function $S(x)$ of a Lagrangian manifold is the function such that if (x,p) is on the Lagrangian manifold, then $p = \partial S(x)/\partial x$; it is generally multivalued, and its branches will be denoted by the index r, as in $S_r(x)$.

A so-called "complete solution" of the Hamilton–Jacobi equation[21] is an f-parameter family of solutions, $S = S(x,a)$, with $a = (a_1, \ldots, a_f)$. We will regard $S(x,a)$ as the generating function of a canonical transformation, in which the new momenta (they could equally well be new coordinates) are $A_1(x,p), \ldots, A_f(x,p)$. Here we use the capital letter A for the new momenta, regarded as functions of (x,p), and the lower case letter a for the values of these functions. We will denote the generalized coordinates conjugate to A by $\alpha = (\alpha_1, \ldots, \alpha_f)$, so that $\alpha_i = \partial S(x,a)/a_i$. Then for each value of a, the Lagrangian manifold $p = p(x,a) = \partial S(x,a)/\partial x$ is the simultaneous contour surface, $A_i(x,p) = a_i, i = 1, \ldots, f$, of the new momenta; and the f-parameter family of Lagrangian manifolds arising by varying the a's is a foliation of phase space into Lagrangian manifolds. The α's serve as coordinates on the Lagrangian manifolds. Since the energy E is constant on each of these Lagrangian manifolds, it is a function of a, i.e., $E = E(a)$. Sometimes it is convenient to choose one of the new momenta to be equal to the Hamiltonian itself, so that E is equal to one of the a's; this is not necessary, however.

We also note that any particular function $S(x)$ obtained from $S(x,a)$ by fixing a, i.e., the generating function of a particular Lagrangian manifold, is a simultaneous solution of the f Hamilton–Jacobi equations,

$$A_i\left(x, \frac{\partial S}{\partial x}\right) = a_i, \quad i = 1, \ldots, f. \tag{2.3}$$

This is almost obvious; if (x,p) is a point on the Lagrangian manifold labeled by a, then $A(x,p) = a$. Thus, when we think we are solving the single Hamilton–Jacobi equation (2.1), we are actually solving f simultaneous Hamilton–Jacobi equations, one for each of a set of f classically commuting observables $A_1, \ldots, A_f$. We will refer to such a set of classical observables as a *complete* set. Conversely, if we ask for the general solution of the simultaneous Hamilton–Jacobi equations (2.3), we find that it is $S(x,a)$, unique to within an additive constant. The constant can be regarded as a convention for the choice of initial point for the contour integral of $p\,dx$ giving $S(x,a)$.

Even when we have only an isolated, particular solution $S(x)$ of the Hamilton–Jacobi equation (2.1), it can always be imbedded in an f-parameter family of solutions. We will actually do this below when we consider the Gutzwiller trace formula for the density of states. Therefore, we can say that every solution of the Hamilton–Jacobi equation is represented by some complete set of commuting classical observables; and, by extension, the same applies to every wave function of semiclassical interest. In particular, it applies to the semiclassical propagator and Green's function; it also applies, as we shall show, to Gutzwiller's formula for the density of states and to Berry's formula for scars of Wigner functions in phase space. The identification of the complete set of commuting classical observables associated with a semiclassical wave function is an important component of its semiclassical interpretation; it is, however, a component which has hitherto been missing in trace formulas.

Let us now suppose that some definite, complete solution $S(x,a)$ of the Hamilton–Jacobi equation has been found, and let us ask for the most general solution of the amplitude transport equation consistent with this $S(x,a)$. The amplitude transport equation is a continuity equation involving the density $\rho(x) = \Omega(x)^2$, regarded as a density in configuration space; one can equally well work with a density $\sigma(\alpha)$ on the Lagrangian manifold, given by

$$\sigma(\alpha) = \rho(x)\left|\det\frac{\partial x}{\partial \alpha}\right|. \tag{2.4}$$

The general solution of the amplitude transport equation is conveniently represented in terms of an initial density on some $(f-1)$-dimensional initial value surface, from which the density is transported along orbits. The initial value surface can be regarded as being in configuration space or in the Lagrangian manifold, depending on whether one wishes to work with $\rho(x)$ or $\sigma(\alpha)$. By choosing an initial density and transporting it along orbits for each Lagrangian manifold, we obtain an f-parameter family of solutions $\Omega(x,a)$ of the amplitude transport equation.

Therefore, the general solution $\Omega(x,a)$ of the amplitude transport equation would seem to involve the selection of an arbitrary initial density, one for each Lagrangian manifold. Indeed, if the Lagrangian manifolds are topologically trivial, i.e., homeomorphic to $\mathbb{R}^f$, as in a scattering problem, then this conclusion is correct. But more generally, Lagrangian manifolds often have the topology of $\mathbb{R}^{f_1}\times(S^1)^{f_2}$, with $f_1 + f_2 = f$, i.e., f_1 lines crossed with f_2 circles.[11] In the extreme case $f = f_2$, the Lagrangian manifold is an f torus. Therefore, it may happen that an orbit leaving the $(f-1)$-dimensional initial surface on a Lagrangian manifold will

return again to this initial surface, usually at a different point from where it left. Continuity then demands that the transported value of the density be equal to the initial density at a point on the initial surface where an orbit returns. In this way, restrictions may be placed on the choice of initial density when the Lagrangian manifolds are topologically nontrivial. In the extreme case that the Lagrangian manifold is an f-torus supporting ergodic orbits, the initial density (and therefore the density everywhere else) is determined to within a multiplicative constant; it is simply $\sigma(\alpha) = \text{const}$.

In this paper we will mostly be interested in the opposite extreme of $f_1 = f$, i.e., topologically trivial Lagrangian manifolds, so our amplitude transport equation will have many possible solutions, even for a given $S(x,a)$. For the same reason, however, we will not need to worry about quantization, since our topologically trivial Lagrangian manifolds will automatically support wave functions that are single valued on the manifold. In other words, for the applications we will consider, the quantities a will be allowed to take on continuous values.

Even when many solutions of the amplitude transport equation exist, there is one that especially stands out, namely,

$$\Omega(x,a) = \left|\det \frac{\partial^2 S(x,a)}{\partial x\,\partial a}\right|^{1/2} = \left|\det \frac{\partial \alpha}{\partial x}\right|^{1/2}. \tag{2.5}$$

Expressed in terms of the density on the Lagrangian manifold, this is the solution $\sigma(\alpha) = 1$. A unique feature of this solution, not shared by other solutions of Eq. (2.2), is that it is actually a simultaneous solution of f amplitude transport equations,

$$\frac{\partial}{\partial x}\left[\Omega(x,a)^2 \frac{\partial A_i(x,p(x,a))}{\partial p}\right] = 0, \quad i = 1, \ldots, f, \tag{2.6}$$

where $p(x,a) = \partial S(x,a)/\partial x$, i.e., one for each of the commuting observables A_i; $\Omega(x,a)$ satisfies these equations because each of the A's, regarded as a Hamiltonian, generates a flow which is a simple displacement in the corresponding α, and a density $\sigma(\alpha)$ which is constant is obviously invariant under these flows. Conversely, if we were to seek a simultaneous solution of Eq. (2.6), the solution is given by Eq. (2.5), and is unique to within a multiplicative constant.

In many applications the complete set of commuting classical constants of motion $A_1(x,p), \ldots, A_f(x,p)$ which emerge from a complete solution of the Hamilton–Jacobi equation are the classical counterparts of a set of commuting quantal constants of motion, $\hat{A}_1, \ldots, \hat{A}_f$. Since these quantal observables commute with $\hat{H}$ and with each other, they possess simultaneous eigenstates that are also eigenstates of the Hamiltonian. We shall denote one of these eigenstates by $|a\rangle$, with $a = (a_1, \ldots, a_f)$ now interpreted as a vector of eigenvalues. As a result, $\psi(x) = \langle x|a\rangle$ is an eigenfunction of the Hamiltonian.

Therefore, we might consider an indirect approach to finding the eigenfunctions of the Hamiltonian, in which we first seek the simultaneous eigenfunctions of the $\hat{A}$'s. If we do this by semiclassical means, we are led to the f simultaneous Hamilton–Jacobi equations (2.3) and the f simultaneous amplitude transport equations (2.6) As we have seen, the solutions of these equations give a wave function $\psi(x)$ which is unique to within an overall multiplicative constant; this constant, whose magnitude comes from the multiplicative constant for $\Omega(x,a)$ and whose phase comes from the additive constant for $S(x,a)$, can be different for different Lagrangian manifolds, i.e., it is a function of a. Its magnitude can be determined by demanding orthonormality, which we apply in the continuum sense, i.e.,

$$\langle a|a'\rangle = \int dx \langle a|x\rangle\langle x|a'\rangle = \delta(a - a'). \tag{2.7}$$

This is equivalent to demanding that $\langle x|a\rangle$ be the component of a unitary transformation matrix, taking us from the x representation to the a representation. The remaining arbitrary phase factor can be regarded as a phase convention for the states $|a\rangle$.

By evaluating the integral of Eq. (2.7) by the stationary phase approximation, we obtain the final expression for the simultaneous eigenfunction of the $\hat{A}$'s:

$$\langle x|a\rangle = \frac{1}{(2\pi i\hbar)^{f/2}} \sum_r \left|\det \frac{\partial^2 S}{\partial x\,\partial a}\right|^{1/2} \times \exp\left[\frac{i}{\hbar} S(x,a) - i\mu\frac{\pi}{2}\right]. \tag{2.8}$$

Here we have introduced the usual Maslov index μ to specify the proper phase shifts between branches r; both S and μ depend on r. We have also introduced for convenience an overall phase factor of $e^{-if\pi/4}$; there remains an additional arbitrary phase which is a function of a. Apart from this, the answer is unique. Of course, this same result will be obtained whenever we seek a simultaneous eigenfunction of a set of f commuting $\hat{A}$'s, whether or not they also commute with $\hat{H}$.

In this paper we will not be much concerned with Maslov indices, so we will simply write μ for them wherever they occur, with no implication that the different μ's are equal. The Maslov index in the Gutzwiller trace formula is considered as a separate issue in papers by Robbins[22] and by Creagh, Robbins, and Littlejohn.[23]

The realization that the requirement of unitarity leads to an essentially unique determination of the amplitude was evidently first made by Miller,[10] so we will refer to Eq. (2.8) as Miller's form of the semiclassical matrix element $\langle x|a\rangle$. Actually, Miller's results were more general than Eq. (2.8), for he gave formulas for the unitary matrix elements $\langle b|a\rangle$, connecting any two representations. Although the modifications required for Miller's general result are easy [one simply replaces x wherever it occurs in Eq. (2.8) by b], nevertheless the implications are far reaching, for Miller's results demonstrate a kind of covariance of semiclassical theory under canonical transformations, exactly mirrored by the covariance of quantum mechanics under unitary transformations. In spirit, Miller's results are fundamentally geometrical. In this paper, there will be special emphasis placed on the property of unitarity in examining the structure of trace formulas, and it will be seen to play a role in unusual contexts.

We will later see examples of complete sets of classically commuting observables, $A_1(x,p), \ldots, A_f(x,p)$, which are either multivalued or discontinuous in phase space. These are the functions one obtains in classical mechanics by attempt-

ing to create constants of motion by demanding that the functions be constant along orbits; except in special circumstances, the contour surfaces of these constants do not have simple imbeddings in phase space, but rather have complicated self-intersections. Such classical observables do not have any clear quantal counterparts, so it is really not meaningful to talk about simultaneous eigenfunctions $\langle x|a\rangle$. Nevertheless, it is possible to construct formal semiclassical expressions of the form of Eq. (2.8), based on the Lagrangian manifolds associated with such classical observables, and we will find it convenient to refer to these expressions with the same kind of bra-ket notation as the $\langle x|a\rangle$ occurring in Eq. (2.8). This notation is especially convenient for discussing Green's functions, as we shall do in Sec. III.

Although the wave function $\psi(x) = \langle x|a\rangle$ of Eq. (2.8) can sometimes represent an eigenfunction of $\hat{H}$, not every eigenfunction of $\hat{H}$ has the form of Eq. (2.8), even in the formal sense of the preceding paragraph. This is in spite of the fact that solving the original Hamilton–Jacobi equation (2.1) always leads to some complete set of A's. The reason is that the trivial solution of the original amplitude transport equation (2.2), given by Eq. (2.5), is not generally the only solution. To find the general solution of Eq. (2.2), let us write $\rho(x) = |\det(\partial\alpha/\partial x)|$ for the square of the trivial amplitude shown in Eq. (2.5). This $\rho(x)$ satisfies the amplitude transport equation in the form

$$\frac{\partial}{\partial x}[\rho(x)v(x)] = 0, \tag{2.9}$$

where the velocity field $v(x)$ is given by

$$v(x) = \frac{\partial H(x,p(x))}{\partial p}, \tag{2.10}$$

with $p(x) = \partial S(x)/\partial x$. (Here we work with a single Lagrangian manifold, and suppress the dependence on a). We now let $\rho'(x)$ be any other solution of Eq. (2.9), and write $\rho'(x) = g(x)\rho(x)$. It follows that

$$v(x)\frac{\partial g(x)}{\partial x} = 0, \tag{2.11}$$

i.e., that g is constant along orbits in configuration space. Choosing some such function $g(x)$, we can convert it into a function $G(x,p)$, defined for (x,p) points on the Lagrangian manifold in question by setting $G(x,p(x)) = g(x)$; the resulting $G(x,p)$ is then constant along orbits on the Lagrangian manifold in phase space. Finally, by carrying out this procedure for the whole family of Lagrangian manifolds, we obtain a function $G(x,p)$ which is defined over a whole finite region of phase space, and which is a constant of motion. We conclude that the most general semiclassical eigenfunction of $\hat{H}$ has the form

$$\psi(x) = \sum_r F(x,p_r(x,a))T_r, \tag{2.12}$$

where T_r is the rth term in Eq. (2.8), where $F(x,p)$ is a constant of motion (the square root of G), and where $p_r = \partial S_r(x,a)/\partial x$.

Notice that the constant of motion in F in Eq. (2.12) need not commute with the A's. Indeed, if it does commute with the A's, then its value $F(x,p)$ on the Lagrangian manifold $A = a$ is simply a function of the a's, and F can be absorbed into the normalization constant for $\langle x|a\rangle$. This would return us to the Miller form of the matrix element, Eq. (2.8). Therefore, F in Eq. (2.12) provides a nontrivial modification to the Miller matrix element only when it does not commute with the A's. One consequence of this is that in one degree of freedom, the semiclassical energy eigenfunctions are uniquely determined by the family of Lagrangian manifolds alone (without specifying the form of the Hamiltonian), because all constants of motion commute with H. But in higher degrees of freedom, there may be more than one way to associate wave functions with families of Lagrangian manifolds. This can happen even in bound state problems, for which the Lagrangian manifolds are tori, if the motion is not ergodic on the tori in a finite volume of phase space. Such nonergodic motion occurs in so-called degenerate classical systems, such as the two-dimensional isotropic harmonic oscillator or the hydrogen atom, in which phase space can be foliated into invariant tori in more than one way.

B. Poisson bracket form for the amplitude

Let us return to Eq. (2.8), and seek an expression for the unitary matrix element $\langle a|b\rangle$, where $\hat{B} = (\hat{B}_1, \ldots, \hat{B}_f)$ is a collection of f new operators that commute with one another (but not necessarily with the $\hat{A}$'s), with eigenvalues $b = (b_1, \ldots, b_f)$ and classical counterparts $B_1(x,p), \ldots, B_f(x,p)$. We do this by writing down the semiclassical formula for $\langle x|b\rangle$, analogous to Eq. (2.8), and by applying the stationary phase approximation to the integral $\int\langle a|x\rangle\langle x|b\rangle$. It is now necessary to distinguish the two generating functions, call them $S_A(x,a)$ and $S_B(x,b)$, producing (with definite values of a and b) two distinct Lagrangian manifolds $p = p_A(x,a)$ and $p = p_B(x,b)$ in phase space. We will call these the A manifold and the B manifold. We will also denote the generalized coordinates conjugate to the B's by $\beta = (\beta_1, \ldots, \beta_f)$, which we use as coordinates on the B manifold.

The computation of the integral is straightforward, except for the amplitude determinant, the reciprocal of which we write as

$$\det\left(\frac{\partial^2 S_A}{\partial x\,\partial a}\right)^{-1}\det\left(\frac{\partial^2 S_B}{\partial x\,\partial b}\right)^{-1}\det\left(\frac{\partial^2 S_A}{\partial x\,\partial x} - \frac{\partial^2 S_B}{\partial x\,\partial x}\right). \tag{2.13}$$

This, however, can be cast into the form

$$\det_{kl}\left[\left(\frac{\partial p_i}{\partial x_j}\right)_b\left(\frac{\partial a_k}{\partial p_j}\right)_x\left(\frac{\partial b_l}{\partial p_i}\right)_x - \left(\frac{\partial p_i}{\partial x_j}\right)_a\left(\frac{\partial a_k}{\partial p_i}\right)_x\left(\frac{\partial b_l}{\partial p_j}\right)_x\right], \tag{2.14}$$

where the quantities being held fixed in the derivatives are explicitly indicated. This in turn can be reduced to

$$\det_{kl}\left[\left(\frac{\partial a_k}{\partial x_j}\right)_p\left(\frac{\partial b_l}{\partial p_j}\right)_x - \left(\frac{\partial a_k}{\partial p_j}\right)_x\left(\frac{\partial b_l}{\partial x_j}\right)_p\right] = \det_{kl}\{A_k,B_l\}, \tag{2.15}$$

where the curly brackets are the usual Poisson bracket. We will simply write this final determinant as $\det\{A,B\}$. Altogether, we have

$$\langle a|b\rangle = \frac{1}{(2\pi i\hbar)^{f/2}}\sum_r \frac{\exp\{(i/\hbar)[S_B(x,b) - S_A(x,a)] - i\mu(\pi/2)\}}{|\det\{A,B\}|^{1/2}}. \tag{2.16}$$

The novel element in this formula is the expression of the amplitude determinant in terms of Poisson brackets.

The branches of the sum in Eq. (2.16) are the intersections of the A manifold with the B manifold, which for now we assume to take place at isolated points; this is the generic situation, since two f-dimensional manifolds in $2f$-dimensional space usually intersect in zero-dimensional points. Therefore, each intersection has an (x,p) value, giving us the x coordinate at which to evaluate the actions $S_A(x,a)$ and $S_B(x,b)$, and the (x,p) coordinates at which to evaluate the Poisson brackets. This situation is illustrated in Fig. 1.

Since the left side of Eq. (2.16) obviously does not depend on any special features of the quantal x representation, it is satisfying that the amplitude determinant on the right appears in a form that is manifestly a phase space invariant, i.e., independent of any special properties of the classical (x,p) coordinates. In this respect, this version of the amplitude determinant is an improvement over that shown in Eq. (2.8). It also has the practical consequence of allowing us to compute the amplitude in any canonical coordinates, which sometimes simplifies calculations.

As for the phase on the right side of Eq. (2.16), we cannot expect it to be a phase space invariant, because it depends on the phase conventions for the states $|a\rangle$ and $|b\rangle$. If our Lagrangian manifolds intersect in more than one point, however, as illustrated in Fig. 2, then the *relative* phase $(S_{B2} - S_{A2}) - (S_{B1} - S_{A1})$ can be expected to be a phase space invariant, because it is not affected by any overall phase factor. Indeed, it is easy to see that the relative phase is simply the symplectic area enclosed between the two intersections. We may note that in many degrees of freedom, where the Lagrangian manifolds are more than one dimensional, the symplectic area is independent of the path chosen to form the loop.

The Maslov index in Eq. (2.16) can also be given an invariant interpretation in terms of the geometry of the Lagrangian manifolds; this is a little more subtle, and we will not go into it here.

Equation (2.16) shows that caustics occur when $\det\{A,B\} = 0$ at a point of intersection of the A manifold and B manifold. The geometrical meaning of this is that the two manifolds are tangent at the point of contact, as illustrated in Fig. 3. In f degrees of freedom, the caustic can be of any order from 1 to f, the order being determined by the corank of the matrix $\{A_k,B_l\}$, i.e., the number of linearly independent null eigenvectors it possesses. These null eigenvectors are related in a simple way to the directions in phase space that are simultaneously tangent to the A manifold and the B manifold at their point of intersection. To see this, let c_l be a null eigenvector of the matrix of Poisson brackets, so that

$$\{A_k,B_l\}c_l = 0. \tag{2.17}$$

Now associate this eigenvector with a phase space vector X^μ, defined by

$$X^\mu = c_l \Gamma^{\mu\nu}\frac{\partial B_l}{\partial \xi^\nu}, \tag{2.18}$$

where Greek indices run from 1 to $2f$ and Latin indices run from 1 to f, where $\xi = (x,p)$ is the $2f$ vector of phase space coordinates, and where Γ is the usual cosymplectic form. For given l, the phase space vector $\Gamma^{\mu\nu}(\partial B_l/\partial \xi^\nu)$ is the flow vector in phase space arising from treating B_l as a Hamiltonian; it represents simply a displacement in the β_l coordinate on the B manifold, and so is tangent to the B manifold. Therefore, X^μ, which is a linear combination of these vec-

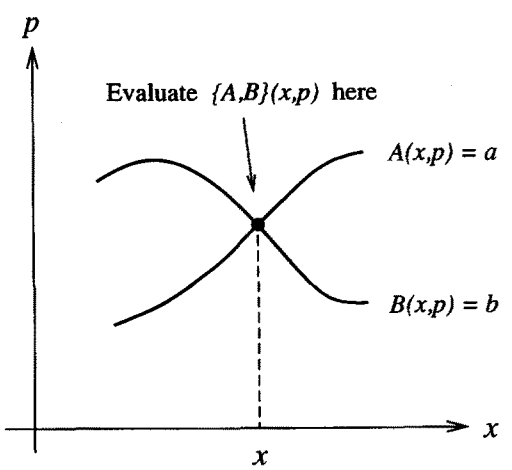

FIG. 1. The semiclassical matrix element $\langle a|b\rangle$ is expressed in terms of the intersections of the Lagrangian manifolds $A(x,p) = a$ and $B(x,p) = b$. The Poisson brackets of the amplitude determinant are evaluated at the intersection points, and the actions in the exponent are evaluated at the x coordinates of the intersection points.

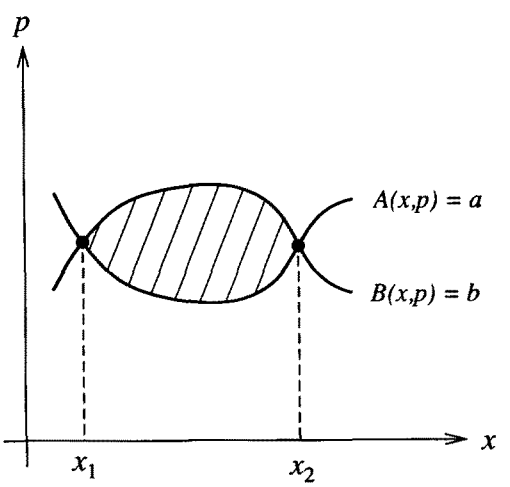

FIG. 2. The relative phase between branches $(S_B(x_2,b) - S_A(x_2,a)) - (S_B(x_1,b) - S_A(x_1,a))$, is a phase space invariant; it is the symplectic area enclosed by the two intersections of the two Lagrangian manifolds.

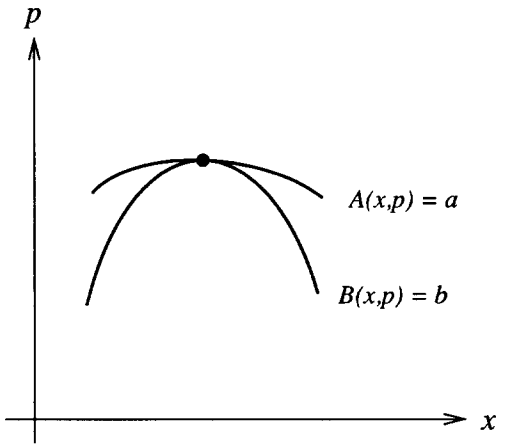

FIG. 3. Caustics of the matrix element $\langle a|b\rangle$ occur when the A manifold and B manifold are tangent at their point of intersection.

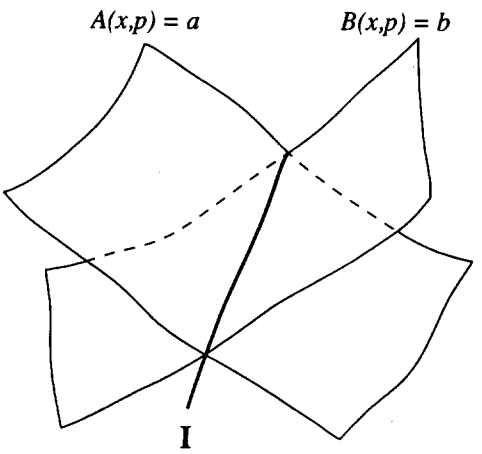

FIG. 4. A case of interest is where the A manifold and B manifold intersect in a region I of dimensionality $f_2 > 0$. In the Gutzwiller trace formula, I is the periodic orbit. This diagram is misleading, in the sense that it suggests that I is stable under small perturbations of the manifolds.

tors, is also tangent to the B manifold. However, X^μ is also tangent to the A manifold, since by Eq. (2.17) we have

$$X^\mu \frac{\partial A_k}{\partial \xi^\mu} = c_l\{A_k,B_l\} = 0. \tag{2.19}$$

Thus X^μ represents a direction in phase space in which the A manifold and B manifold are tangent to one another.

If $\det\{A,B\}$ should vanish at an intersection point, the most typical case would be that the intersection point would be isolated and $\{A_k,B_l\}$ would have a single null eigenvector there. Then the right side of Eq. (2.16) for the matrix element $\langle a|b\rangle$ would be replaced by an expression involving an Airy function. It would be straightforward to analyze this case, and to build the required expression out of elements whose invariant meaning in phase space is manifest.

For the applications we shall consider, however, the A and B manifolds intersect, not at an isolated point, but over a whole region of dimensionality $f_2 > 0$, with $f = f_1 + f_2$. We will call this f_2-dimensional intersection I, as illustrated in Fig. 4, so that $\{A_k,B_l\}$ has rank f_1 everywhere on I. For example, we will show in Sec. IV that in the Gutzwiller trace formula for nonintegrable systems, I is identified with a periodic orbit, in which case $f_2 = 1$.

Figure 4 is misleading in one respect, namely, that it suggests that the intersection I changes continuously under small changes in the A and B manifolds. Actually, such an intersection with dimensionality $f_2 > 0$ is unstable, and breaks up into isolated points under most perturbations of the manifolds. Indeed, the very existence of such nontrivial intersections seems to be related to symmetry; for example, in the standard Gutzwiller trace formula, it is related to conservation of energy.

In the cases of interest to us, the A and B manifolds intersect in an f_2-dimensional intersection I because f_2 of the A's are identical with f_2 of the B's (or because coordinate transformations of the A's among themselves and of the B's among themselves can bring this situation about). This is fortunate, because in this case the computation of the semiclassical matrix element $\langle a|b\rangle$ is easier than in the general case.

To carry out this computation, we write $A = (A_1,A_2)$, $B = (B_1,B_2)$, where the first and second members contain f_1 and f_2 functions, respectively, and where $B_2 = A_2$. With no loss of generality we also set $\alpha_2 = \beta_2$. In this way, α_2 or β_2 can be taken as coordinates on I, whereas both α_1 and β_1 are constant on I.

This coordinate system simplifies the calculation of the matrix element $\langle a|b\rangle$, which we carry out in the α representation, performing the stationary phase approximation on the integral $\int d\alpha \langle a|\alpha\rangle\langle\alpha|b\rangle$. The actions in the α representation are the integrals of $a\,d\alpha$ along the respective manifolds; for the A manifold we have $S_A(\alpha,a) = \alpha a$, so $\det(\partial^2 S_A/\partial\alpha\,\partial a) = 1$. Therefore, the integral we must evaluate is

$$\int d\alpha \left|\det \frac{\partial^2 S_B(\alpha,b)}{\partial\alpha\,\partial b}\right|^{1/2} \exp\left\{\frac{i}{\hbar}[S_B(\alpha,b) - \alpha a]\right\}. \tag{2.20}$$

Let us write simply $S(\alpha)$ for the total action $S_B(\alpha,b) - S_A(\alpha,a)$ in the exponent; it is constant on I, because as we move along I, the increment in S_B cancels that in S_A. Therefore, in expanding $S(\alpha)$ to second order about a point on I we have only the α_1 derivatives to take. For convenience we let I be specified by $\alpha_1 = 0$, so the expansion of $S(\alpha)$ is

$$S(\alpha) = S(I) + \frac{1}{2}\alpha_1 \frac{\partial^2 S}{\partial\alpha_1\,\partial\alpha_1}\alpha_1, \tag{2.21}$$

where the linear terms cancel because I is the stationary phase set.

As for the amplitude prefactor, we note first that

$$\frac{\partial^2 S_B(\alpha,b)}{\partial\alpha_2\,\partial b_1} = \left(\frac{\partial\beta_1}{\partial\alpha_2}\right)_b = 0, \tag{2.22}$$

because to vary α_2 while holding b fixed is to move along I, where β_1 is constant. We also have

$$\frac{\partial^2 S_B(\alpha,b)}{\partial\alpha_2\,\partial b_2} = \left(\frac{\partial\beta_2}{\partial\alpha_2}\right)_b = 1, \tag{2.23}$$

since we have $\alpha_2 = \beta_2$ on I. Therefore,

$$\det\left(\frac{\partial^2 S_B(\alpha,b)}{\partial\alpha\,\partial b}\right) = \det\left(\frac{\partial^2 S_B(\alpha,b)}{\partial\alpha_1\,\partial b_1}\right). \tag{2.24}$$

Now carrying out the integral over α_1, we obtain an amplitude determinant whose reciprocal can be written,

$$\begin{aligned}
&\det\left(\frac{\partial b_1}{\partial a_1}\right)_{\alpha_1} \det\left(\frac{\partial a_1}{\partial \alpha_1}\right)_{b_1} \\
&\quad = \det\left[-\left(\frac{\partial b_1}{\partial \alpha_1}\right)_{a_1}\right] \\
&\quad = \det\left[\left(\frac{\partial A_1}{\partial \alpha_1}\right)_{a_1}\left(\frac{\partial B_1}{\partial a_1}\right)_{\alpha_1} - \left(\frac{\partial A_1}{\partial a_1}\right)_{\alpha_1}\left(\frac{\partial B_1}{\partial \alpha_1}\right)_{a_1}\right] \\
&\quad = \det\{A_1,B_1\}_1,
\end{aligned} \tag{2.25}$$

where in all the derivatives shown, $\alpha_2 = \beta_2$ and $a_2 = b_2$ are held fixed, and where the 1 subscript on the Poisson bracket indicates that the bracket is computed with respect to the f_1 canonical pairs (α_1,a_1) only. The right side is a determinant of an $f_1 \times f_1$ matrix of Poisson brackets. The 1 subscript on the Poisson bracket can be dropped, however, because the terms coming from the (α_2,a_2) canonical pairs vanish, due to the vanishing of the derivatives $\partial A_1/\partial\alpha_2$ and $\partial A_1/\partial a_2$. Therefore, the Poisson brackets $\{A_{1k},B_{1l}\}$ are standard Poisson brackets on the full phase space, and can be computed in any canonical coordinates.

These Poisson brackets are also independent of α_2, and can therefore be taken out of the α_2 integral. We show this by taking the α_2 derivative, which can be written as a Poisson bracket and transformed by using the Jacobi identity:

$$\begin{aligned}
&\frac{\partial}{\partial\alpha_{2i}}\{A_{1k},B_{1l}\} \\
&\quad = \{\{A_{1k},B_{1l}\},A_{2i}\} \\
&\quad = -\{\{B_{1l},A_{2i}\},A_{1k}\} - \{\{A_{2i},A_{1k}\},B_{1l}\} = 0.
\end{aligned} \tag{2.26}$$

The final equality follows because $\{A_{2i},A_{1k}\} = 0$, since all the A's commute with one another, and because $\{B_{1l},A_{2i}\} = \{B_{1l},B_{2i}\} = 0$, since all the B's commute.

The final result is

$$\begin{aligned}
\langle a|b\rangle = &\frac{1}{(2\pi i\hbar)^{(f+f_2)/2}} \\
&\times\sum_r \frac{\exp\{(i/\hbar)[S_B(I) - S_A(I)] - i\mu(\pi/2)\}}{|\det\{A_1,B_1\}|^{1/2}} \\
&\times\int_I d\alpha_2.
\end{aligned} \tag{2.27}$$

The integral shown is just the α_2 volume of I. In Sec. IV we will show that the Gutzwiller formula for the density of states of a classically nonintegrable system is a matrix element of this kind.

III. SEMICLASSICAL STRUCTURE OF THE PROPAGATOR AND GREEN'S FUNCTION

In this section we analyze the role of the propagator and the Green's function in Gutzwiller's derivation of his formula for the density of states. We do this from the standpoint of Maslov's version of WKB theory, and from that of the general properties of semiclassical wave functions and matrix elements as discussed in Sec. II.

By the propagator we mean the function $K(x,t;x',t') = \langle x|U(t,t')|x\rangle\Theta(t-t')$, where $U(t,t')$ is the unitary time evolution operator of a quantal system, taking the system from initial time t' to final time t, and where Θ is the unit step function. If the Hamiltonian is time independent, we will set $t' = 0$ and write simply $K(x,x',t)$ for the propagator. The propagator is sometimes also called the "time-dependent Green's function," but we will not use that term; instead, we will reserve the term "Green's function" for the energy-dependent Green's function $G(x,x',E)$, defined in Eq. (3.12) below as the one-sided Fourier transform in time of the propagator.

The original motivation for the work presented in this section was to simplify Gutzwiller's derivation of the trace formula, or at least to see what general principles were operative in it. As a result, we do not in this section take an approach to the derivation of the trace formula which is substantially different from Gutzwiller's, nor do we carry the derivation through to completion. The reader who is interested in a new perspective on this derivation may skip to Sec. IV, with little loss in continuity.

Nevertheless, several points of interest did emerge from the analysis of Gutzwiller's derivation, and it seemed worthwhile to put them down here. For example, we will show in this section that both the propagator and the Green's function are represented in the extended phase space by the same Lagrangian manifold, showing that both suffer in exactly the same way from the development of "whorls and tendrils" (in the terminology of Berry *et al.*),[24] i.e., complicated convolutions. We will also show how several of Gutzwiller's amplitude determinants, which he transformed into one another by difficult calculations, can be understood as special cases of Miller's formula, Eq. (2.8), or its generalizations. We will also point out that the semiclassical formula for the Green's function bears a close relation to a semiclassical version of a "propagator" for the surface of section evolution. This fact is almost immediately obvious, once it is realized that one of Gutzwiller's formulas is actually more general than his original analysis indicated, but nevertheless it is an interesting and potentially important fact.

A. The propagator

The semiclassical theory of the propagator is an old subject, going back at least to Van Vleck.[25] Its main result is the Van Vleck formula:

$$\begin{aligned}
K(x,t;x',t') = &\frac{1}{(2\pi i\hbar)^{f/2}}\sum_r\left|\det\frac{\partial^2 R}{\partial x\,\partial x'}\right|^{1/2} \\
&\times\exp\left[\frac{i}{\hbar}R(x,t;x',t') - i\mu\frac{\pi}{2}\right],
\end{aligned} \tag{3.1}$$

where R is Hamilton's principal function, i.e., the integral of $p\,dx - H\,dt$ along an orbit connecting configuration space points x' and x at initial and final times t' and t, respectively; where r is the index of the sum over the branches of R, i.e., the distinct orbits satisfying the given end-point conditions; and where μ is the Maslov index. Perhaps the most physically appealing way of deriving this formula is by applying the

stationary phase approximation to the Feynman path integral.[26] However, it can also be derived within time-dependent WKB theory, as a special case of the initial value problem. This is the approach taken by Van Vleck himself, and it has the advantage that it applies to Hamiltonians or other evolution operators of quite general functional form, i.e., operators for which the formulation of the Feynman path integral is difficult.

Much has been written about the Van Vleck formula and its derivation, either from the Feynman path integral or via time-dependent WKB theory. For example, a good deal of the work by Maslov[5] concerns time-dependent WKB theory, of which the most important application is the Van Vleck formula. A review and summary of earlier literature on the Van Vleck formula may be found in Berry and Mount,[27] and the subject has been covered again more recently by Ozorio de Almeida.[8] Therefore, here we will simply summarize some of the principal features of the derivation of the Van Vleck formula from time-dependent WKB theory, and then take up a slightly different approach, one based on time-independent WKB theory.

In applying time-dependent WKB theory to the propagator, one begins with the observation that $K(x,t;x',t')$, regarded as a wave function in the (x,t) variables, satisfies the inhomogeneous, time-dependent Schrödinger equation,

$$\left(\hat{H} - i\hbar\frac{\partial}{\partial t}\right)K(x,t;x',t') = -i\hbar\delta(x-x')\delta(t-t'). \tag{3.2}$$

It follows from this that if one takes an initial wave function $\psi(x,t') = \delta(x-x')$ and solves for $\psi(x,t)$ for $t > t'$, then the exact solution is $\psi(x,t) = K(x,t;x',t')$; and it turns out that the WKB approximation to this solution is the Van Vleck formula. One slight difficulty in applying WKB theory is the singularity of the initial conditions, due to the severe caustic (of order f) which $\psi(x,t)$ develops as $t \to t'$. A number of authors have dealt with this problem by making a special analysis of the Schrödinger equation for small times; this leads to the conclusion that the behavior of the short-time propagator is dominated by the kinetic energy term in the Hamiltonian, so that the free particle results can be used. This approach has the drawback that it applies only to a restricted class of Hamiltonians. A better approach is to follow Ozorio de Almeida,[8] who applies Maslov's method[5] of switching to the momentum representation near caustics. That is, one computes the function $\tilde{K}(p,t;x',t') = \langle p|U(t,t')|x'\rangle$ for t near t', in the WKB approximation. (The propagator has been transformed to the momentum representation in the final variable, with the initial variable being treated as a parameter.) Not only is this function well behaved, but one can also get explicit formulas for the short-time amplitude and action, in terms of an arbitrary Hamiltonian. In this way the initial conditions can be satisfied, and the Van Vleck formula derived; one result is that the form shown in Eq. (3.1) is valid for general Hamiltonians.

In Sec. II above we considered only time-independent WKB theory, but nevertheless many of the results obtained there apply to the Van Vleck formula. For example, the Van Vleck formula is a special case of Miller's matrix element, Eq. (2.8), as was noted by Miller himself.[10] To show this, we let $\hat{x}$ be the usual position operator in quantum mechanics, and we let $\hat{x}' = U(t,t')\hat{x}U(t,t')^{\dagger}$. We then denote the eigenstates of $\hat{x}$ and $\hat{x}'$ with eigenvalues x and x' by $|\hat{x}(x)\rangle$ and $|\hat{x}'(x')\rangle$, respectively, so that

$$|\hat{x}'(x')\rangle = U(t,t')|\hat{x}(x')\rangle. \tag{3.3}$$

It follows that

$$K(x,t;x',t') = \langle\hat{x}(x)|\hat{x}'(x')\rangle, \tag{3.4}$$

which we might simply write as $\langle x|x'\rangle$, if it were understood that the operators whose eigenvalues are given in the bra and ket are the distinct operators $\hat{x}$ and $\hat{x}'$.

Therefore, on comparing the Van Vleck formula with Eq. (2.8), we see that $\hat{A} = \hat{x}'$ and $a = x'$. The classical counterpart of $\hat{A}$ is $A(x,p,t,t') = X'(x,p,t,t')$, where X' is the initial position along an orbit, considered as a function of the final (x,p) and the two times. The components of X' commute with one another, because the classical transformation from initial conditions (x',p') to final conditions (x,p) is a canonical transformation. The Lagrangian manifold in the (x,p) phase space corresponding to the Van Vleck formula for fixed values of (x',t',t), regarded as parameters, is the set of points in phase space with the given value of initial position x'. Equation (2.8) does not by itself give the overall phase of the Van Vleck formula; this is important, and can be determined by other means. Finally, we note that the amplitude determinant in the Van Vleck formula can be written in terms of Poisson brackets as in Eq. (2.16). The bracket expression is simply $\det\{x,x'\}$; computing this with respect to the (x,p) or (x',p') canonical coordinates, we obtain $\det(\partial x'/\partial p)_x$ or $\det(-\partial x/\partial p')_{x'}$, both of which are identical to $\det(\partial^2 R/\partial x\,\partial x')^{-1}$.

We turn now to an application of time-independent WKB theory to the propagator. Actually, it will be better to talk about "autonomous WKB theory," since the independent variable will no longer be time. This kind of approach to the propagator has been at least mentioned by Arnold;[18] here we will expound on it in somewhat more detail, since it provides some interesting perspectives on both the propagator and Green's function, and since it helps clarify some of Gutzwiller's calculations.

We begin by enlarging the classical phase space with coordinates (x,p) by one degree of freedom, producing an "extended" phase space with coordinates (x,p,t,w). Here, the time t has been demoted from the status of the independent variable to that of a dependent variable, producing an extended configuration space with coordinates (x,t), and the momentum w conjugate to time has been introduced. The extended Hamiltonian is

$$\mathcal{H}(x,p,t,w) = H(x,p,t) + w, \tag{3.5}$$

where $H(x,p,t)$ is the original Hamiltonian. It produces the autonomous (τ independent) equations of motion,

$$\begin{aligned} \frac{dx}{d\tau} &= \frac{\partial H}{\partial p}, & \frac{dp}{d\tau} &= -\frac{\partial H}{\partial x}, \\ \frac{dt}{d\tau} &= 1, & \frac{dw}{d\tau} &= -\frac{\partial H}{\partial t}, \end{aligned} \tag{3.6}$$

where τ is the parameter of the orbits. The extended Hamil-

tonian $\mathscr{H}$ is conserved along the orbits in the extended phase space, and the $(2f+1)$-dimensional surface $\mathscr{H}=\mathscr{E}$ can be regarded as an extended energy shell. The intersection of this energy shell with the surface $t=\text{const}$, say, $t=t'$, can be regarded as an extended surface of section; it is crossed by all orbits, since $dt/d\tau=1$. The extended surface of section is naturally identified with the original phase space (x,p) at the given, constant time. The extended energy shell and surface of section are sketched in Fig. 5.

By replacing the classical $H(x,p,t)$ by its quantal counterpart $\hat{H}$, and by replacing w by the operator $-i\hbar\,\partial/\partial t$, we can also write down an extended Schrödinger equation,

$$\hat{\mathscr{H}}\psi(x,t)=\left(\hat{H}-i\hbar\frac{\partial}{\partial t}\right)\psi(x,t)=\mathscr{E}\psi(x,t), \tag{3.7}$$

in which we seek the extended eigenfunction $\psi(x,t)$ of $\hat{\mathscr{H}}$ with eigenvalue $\mathscr{E}$, defined on the extended configuration space. Notice that every solution $\psi(x,t)$ of the ordinary, time-dependent Schrödinger equation is an eigenfunction of $\hat{\mathscr{H}}$ of eigenvalue $\mathscr{E}=0$, and conversely. Only the value $\mathscr{E}=0$ has physical significance, although it turns out that we must find eigenfunctions for arbitrary values of $\mathscr{E}$ in order to use Miller's form of the amplitude determinant. This equation allows us to regard the solution of the ordinary, time-dependent Schrödinger equation as an eigenvalue–eigenfunction problem in the extended space, and apply autonomous (in this case, τ independent) WKB theory. The eigenfunctions we will find are somewhat like scattering solutions of the usual Schrödinger equation, since they are unbounded in t; thus, the spectrum of $\hat{\mathscr{H}}$ is continuous.

To solve the extended Hamilton–Jacobi equation coming from Eq. (3.7), we require invariant Lagrangian manifolds in the extended phase space, which are conveniently constructed by the surface of section method. Corresponding to the initial conditions inherent in the propagator, we choose a Lagrangian manifold in the (x,p) phase space at time $t=t'$ which is given by $x=x'$ (a vertical line in Fig. 5).

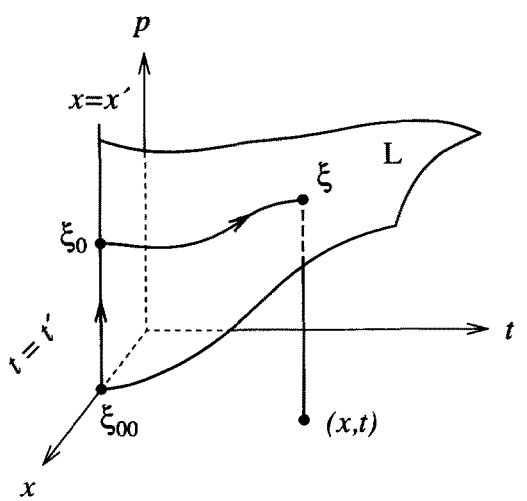

FIG. 5. The extended energy shell $\mathscr{H}=\mathscr{E}$ in the extended phase space, on which the coordinates are (x,p,t). The Lagrangian manifold L is swept out by orbits emanating from the surface $x=x'$ at $t=t'$, which is an initial Lagrangian manifold in the (x,p) phase space. The Lagrangian manifold L supports the semiclassical wave functions for both the propagator and the Green's function, when they are viewed in the extended space.

This initial Lagrangian manifold is allowed to flow into the extended energy shell, sweeping out the $(f+1)$-dimensional Lagrangian manifold L in the figure. This Lagrangian manifold can be promoted into an $(f+1)$-parameter family, thereby foliating the extended phase space, by varying x' and $\mathscr{E}$; in this way we obtain an extended generating function $\mathscr{S}(x,t;x',\mathscr{E})$. On comparing this with the function $S(x,a)$ discussed in Sec. II, we must identify (x,t) with x and $(x',\mathscr{E})$ with a. As in Sec. II, it helps to distinguish functions from values, and to write $(X',\mathscr{H})$ for the commuting classical observables whose values are $(x',\mathscr{E})$ on a given Lagrangian manifold. Here, $X'=X'(x,p,t,t')$ has the same meaning as earlier, i.e., the initial position on an orbit; notice that it is a τ-independent function on the extended phase space (since t is now just another phase space coordinate). Thus the observables $A(x,p)$ of Sec. II are identified with $(X',\mathscr{H})$, both considered to be functions of (x,p,t,w). Notice also that the initial time t' is simply a parameter in this picture; unlike t, it is not a coordinate on the extended phase space.

The extended generating function is easily computed. It is

$$\mathscr{S}(x,t;x',\mathscr{E})=\int p\,dx+w\,dt=\int p\,dx+(\mathscr{E}-H)dt, \tag{3.8}$$

where the fact that $\mathscr{H}=\mathscr{E}$ on the extended energy shell has been used to write $w=\mathscr{E}-H$. The contour of integration is taken from some fixed, initial point on the Lagrangian manifold to a final point ξ, which is the point (or one of the points) on L directly above the observation point (x,t) in the extended configuration space at which we wish to evaluate $\psi(x,t)$. It is convenient to take the initial point to be ξ_{00} in the figure, with coordinates $x=x'$, $p=0$, $t=t'$. The contour of integration could be any contour on L joining ξ_{00} and ξ, but it is convenient to let it run up the initial Lagrangian manifold in the surface of section to the point ξ_0, which is the initial condition for the orbit starting in the surface of section at time t' and terminating on ξ at time t. The contour then continues to the final point ξ by following along the orbit. The integral along the first segment of the contour, from ξ_{00} to ξ_0, vanishes, since $dx=dt=0$ on this segment; therefore, the only contribution is that along the orbit, and $\mathscr{S}$ is easily connected with Hamilton's principal function R. The result is

$$\mathscr{S}(x,t;x',\mathscr{E})=R(x,t;x',t')+\mathscr{E}(t-t'). \tag{3.9}$$

Consider now for a moment a certain exact solution $\psi(x,t)$ of the extended Schrödinger equation (3.7), which is easy to express in terms of the exact propagator K. We choose this wave function to be a simultaneous eigenfunction of $\hat{\mathscr{H}}$ and the operators $\hat{x}'$ used in Eq. (3.3), so that it can be regarded as the unitary matrix element $\langle x,t\,|x',\mathscr{E}\rangle$. Then we have

$$\langle x,t\,|x',\mathscr{E}\rangle=K(x,t;x',t')\exp\left[i\mathscr{E}(t-t')/\hbar\right], \tag{3.10}$$

in which for simplicity we have dropped overall multiplicative constants.

On the other hand, the WKB expression for $\langle x,t\,|x',\mathscr{E}\rangle$ is easy to write down, by using the extended action $\mathscr{S}$ of Eq. (3.9) and Miller's form of the matrix element. As for the

amplitude determinant, we have

$$\det\begin{pmatrix} \frac{\partial^2 \mathscr{S}}{\partial x\,\partial x'} & \frac{\partial^2 \mathscr{S}}{\partial x\,\partial \mathscr{E}} \\ \frac{\partial^2 \mathscr{S}}{\partial t\,\partial x'} & \frac{\partial^2 \mathscr{S}}{\partial t \partial \mathscr{E}} \end{pmatrix} = \det\begin{pmatrix} \frac{\partial^2 R}{\partial x\,\partial x'} & 0 \\ \frac{\partial^2 R}{\partial t\,\partial x'} & 1 \end{pmatrix} = \det \frac{\partial^2 R}{\partial x\,\partial x'}, \tag{3.11}$$

from which the Van Vleck formula immediately follows.

The point of this calculation is to identify the Lagrangian manifold in the extended phase space associated with the propagator, and to find the corresponding generating function. It turns out that the same Lagrangian manifold supports the Green's function, regarded as a wave function in the extended phase space. We now turn to this connection.

B. The Green's function

In the following discussion of the Green's function, we assume that H is time independent, so we set $t' = 0$ and write simply $K(x,x',t)$ for the propagator. Our convention for the Green's function is

$$G(x,x',E) = \frac{1}{i\hbar}\int_0^\infty dt\, e^{iEt/\hbar} K(x,x',t), \tag{3.12}$$

where the convergence of the integral is guaranteed by giving E a positive imaginary part. The Green's function satisfies the inhomogeneous, time-independent Schrödinger equation,

$$(E - H)G(x,x',E) = \delta(x - x'), \tag{3.13}$$

where H acts on the unprimed variables.

The evaluation of the integral of Eq. (3.12) by the stationary phase approximation has been carried out by Gutzwiller,[1] who ignores the contributions coming from the end point $t = 0$. These contributions are important, as they ultimately yield the Fermi–Thomas or averaged density of states. In spite of this, we will also ignore these terms, since we are mainly interested in the phase space structure associated with orbits of nonzero length. We point out, however, that these terms have much in common with the expressions that arise in Keller's geometrical theory of diffraction,[28] which are due to discontinuities in the boundary conditions for a wave field.

Gutzwiller finds the following semiclassical expression for the Green's function:

$$G(x,x',E) = \frac{1}{i\hbar}\frac{1}{(2\pi i\hbar)^{(f-1)/2}} \times \sum_r |D|^{1/2} \exp\left[\frac{i}{\hbar} S(x,x',E) - i\mu\frac{\pi}{2}\right], \tag{3.14}$$

where the sum is taken over all orbits of energy E connecting x' at $t = 0$ and x at a later time; where $S(x,x',E)$ is the reduced action, i.e., the integral of $p\,dx$ along the orbit; and where D is an amplitude determinant, for which Gutzwiller gives two forms. Since the computation of these determinants is a difficult aspect of Gutzwiller's derivation, it is worthwhile to examine their structure from a general standpoint.

Gutzwiller's first form for D is an $(f+1)\times(f+1)$ form,

$$D = \det\begin{pmatrix} \frac{\partial^2 S}{\partial x\,\partial x'} & \frac{\partial^2 S}{\partial x\,\partial E} \\ \frac{\partial^2 S}{\partial E\,\partial x'} & \frac{\partial^2 S}{\partial E^2} \end{pmatrix}, \tag{3.15}$$

and it seems obvious that this must represent some kind of density on an extended space. Indeed it does; we simply transform the extended wave function $\psi(x,t) = \langle x,t\,|x',\mathscr{E}\rangle$ in Eq. (3.10) from the (x,t) representation to the (x,w) representation, i.e., we compute

$$\phi(x,w) = \langle x,w|x',\mathscr{E}\rangle = \int e^{-iwt/\hbar}\langle x,t\,|x',\mathscr{E}\rangle dt, \tag{3.16}$$

again ignoring multiplicative constants. Evaluating this integral by the stationary phase approximation and retaining only the positive time stationary phase points, we obtain the contributions of interest to the semiclassical expression for $G(x,x',E)$, evaluated at $E = \mathscr{E} - w$. On the other hand, we can get the same result from the Lagrangian manifold associated with the wave function $\phi(x,w)$ of Eq. (3.16); since only the representation has changed from Eq. (3.10), but not the commuting operators or the eigenvalues, the manifold is the same as in Fig. 5. In other words, the propagator and the Green's function are represented by the same manifold in the extended phase space.

It is easy to compute the generating function $\tilde{\mathscr{S}}(x,w;x',\mathscr{E})$, where the tilde distinguishes the (x,w) representation from the (x,t) representation that was shown in Eq. (3.9). The result is

$$\tilde{\mathscr{S}}(x,w;x',\mathscr{E}) = \mathscr{S}(x,t;x',\mathscr{E}) - tw = S(x,x',\mathscr{E} - w), \tag{3.17}$$

from which the Miller form of the determinant is easily computed. We find

$$\det\begin{pmatrix} \frac{\partial^2 \mathscr{S}}{\partial x\,\partial x'} & \frac{\partial^2 \mathscr{S}}{\partial x\,\partial \mathscr{E}} \\ \frac{\partial^2 \mathscr{S}}{\partial w\,\partial x'} & \frac{\partial^2 \mathscr{S}}{\partial w\,\partial \mathscr{E}} \end{pmatrix} = \det\begin{pmatrix} \frac{\partial^2 S}{\partial x\,\partial x'} & \frac{\partial^2 S}{\partial x\,\partial E} \\ -\frac{\partial^2 S}{\partial E\,\partial x'} & -\frac{\partial^2 S}{\partial E^2} \end{pmatrix} = -D, \tag{3.18}$$

where we set $E = \mathscr{E} - w$. In this way, Gutzwiller's $(f+1)\times(f+1)$ determinant is seen to be a special case of Miller's determinant in Eq. (2.8), and the Green's function, insofar as its semiclassical structure is concerned, is seen to be a unitary matrix element on the extended space. That is, we are tempted to write

$$\langle x,w|x',\mathscr{E}\rangle = G(x,x',\mathscr{E} - w); \tag{3.19}$$

however, this is only a schematic expression, since the time integral in Eq. (3.12) is only taken from 0 to ∞, and since we have dropped overall multiplicative constants. This expression does, however, capture all the semiclassical features of the positive time stationary phase points, when Eq. (2.8) is

applied to the left side and Eq. (3.14) to the right.

The Lagrangian manifold L of Fig. 5, which supports both the propagator and the Green's function, will in general develop complicated convolutions in the course of time. The formation of convolutions in Lagrangian manifolds has been studied in a detailed analysis by Berry *et al.*[24] These authors find two different morphologies for the convolutions, designated "whorls" and "tendrils," depending on whether the Lagrangian manifold passes through an integrable or chaotic region of phase space, respectively. Their studies concerned area-preserving maps, i.e., systems of one and one half degrees of freedom, but their results are certainly applicable as well to the present discussion of Lagrangian manifolds in the extended phase space. Their analysis reveals a critical time, beyond which the Lagrangian manifolds can no longer predict the details of the exact quantal wave functions, due to the formation of convolutions whose area is smaller than $O(\hbar)$.

One must suppose, therefore, that in a formal sum over an infinite number of branches of the solutions of the Hamilton–Jacobi equation, such as Gutzwiller's formula for the Green's function, Eq. (3.14), only some of the terms can be valid. The valid terms are presumably those corresponding to pieces of the Lagrangian manifold that locally have convolutions on a scale larger than some power of $\hbar$, say, $\hbar^{\nu}$. Here, we will not worry about the precise value of ν, but in all analyses of this sort, including those based on wave packets, ν is of order unity. The valid terms can be thought of as coming from orbits for which the stretching factor (appropriately defined) is less than $\hbar^{-\nu}$; for chaotic systems, this implies a restriction to orbits whose elapsed time t is less than $T=(\nu/\lambda)|\ln\hbar|$, where λ is the Liapunov exponent.

In the case of the Green's function, the valid terms can be selected out and the others rejected if the imaginary part ϵ given to the energy is set to $\hbar/T$, since this is equivalent to multiplying the integrand of Eq. (3.12) (for real E) by $e^{-t/T}$. This is also equivalent to replacing the δ functions in the density of states by Lorentzians of width ϵ, giving a smoothed density of states. Since the average separation between energy levels scales as $\hbar^{-f}$ for f degrees of freedom, we expect the smoothed density of states to be able to resolve energy scales containing a minimum number of individual eigenstates which scales with $\hbar$ as $|\ln\hbar|/\hbar^{f-1}$.

On the basis of these considerations, we do not expect to be able to resolve individual eigenstates in chaotic systems by semiclassical means. This restriction is not really due to the details of the formalism based on Green's functions, but rather to the limitations of any structure based purely upon classical mechanics. For example, the same general kind of limitation occurs in Heller's analysis[29] of wave packets in chaotic systems. These limitations are ultimately based on a combination of the exponential divergence of nearby orbits and the time-energy uncertainty relations, and seem difficult to avoid.

On the other hand, it may be true that some classical structure is still relevant beyond the critical time of order $\ln|\hbar|$. For example, Berry *et al.*[24] found that smoothed wave functions did seem to agree with smoothed, nominal semiclassical predictions, even beyond the critical time; and recent results by Saraceno[30] seem to show the maintenance of certain classical structures in the same regions of time, in the case of the quantized Baker's map.

Voros[31] has also made some comments recently that are relevant in this regard. Voros points out that the knowledge of certain details of the asymptotic (long-time) distribution of the actions of periodic orbits is necessary in order to perform an analytic continuation of Euler product representation of the quantum functional determinant down to the real energy axis, where it would presumably yield the energy eigenvalues. That this process has actually be carried out by Gutzwiller[32] for the anisotropic Kepler problem—not exactly, but with impressive accuracy—suggests that very long orbits retain a meaning, even beyond the length of time for which WKB theory is able to guarantee them a meaning. On the other hand, Voros also points out that quantal corrections, which might in our case be identified with the errors of the semiclassical expansion which grow up and dominate at the critical time of order $\ln|\hbar|$, might invalidate the process of analytical continuation. Thus Guzwiller's results for the anisotropic Kepler problem might be special and not generalizable.

Some of these issues and others as well are discussed in the review article by Berry,[7] which includes a number of observations not found elsewhere. The problem of the long-time behavior of quantum states and its relation to classical structures, if any, is an important one. It is, however, somewhat outside the main thrust of this paper, and so we will now turn to other matters.

C. Second form of the amplitude determinant

Gutzwiller's second form for the amplitude determinant D of the Green's function is an $(f-1)\times(f-1)$ form,

$$D=\frac{1}{\dot{z}'\dot{z}''}\det\left(\frac{\partial^2 S}{\partial y''\,\partial y'}\right), \tag{3.20}$$

where here and in the following discussion we use double primes for a final value such as z'' when it is necessary to distinguish it from a variable value such as z. We continue to use single primes for an initial value such as z'. In Gutzwiller's original derivation, Eq. (3.20) was intended to be applied in the immediate neighborhood of a given periodic orbit, in which z was the distance along the periodic orbit in configuration space, y was a set of $f-1$ coordinates orthogonal to the periodic orbit, and where the initial and final points x' and x'' of the Green's function are set equal to one another in preparation for taking the trace.

Actually, none of these restrictions is necessary, as we may show by retracing the derivation of Eq. (3.20) from Eq. (3.15). To begin, let us consider a point transformation taking us from x to a new set of configuration space coordinates (y,z), in which y is an $(f-1)$ vector of coordinates and z is a single coordinate. We do not necessarily tie these to a periodic orbit, but we do select z to be a convenient independent variable for the evolution in the energy shell $H=E$. (This is the ordinary energy shell in the ordinary phase space, not the extended objects considered previously.) We conceive of this evolution as taking the system from an initial surface of

section $z = z'$ to a final one $z = z''$. For example, z could be one of the original x's, as is common in applications. The energy shell and the z evolution within it are illustrated schematically in Fig. 6.

The Hamiltonian for the z evolution is $F = -p_z$, where F is expressed as a function of (y,p_y,z) by solving $H = E$ for p_z, and is parameterized by E. The z evolution is governed by Hamilton's equations, which are generally nonautonomous (i.e., z dependent),

$$\frac{dy}{dz} = \frac{\partial F}{\partial p_y}, \quad \frac{dp_y}{dz} = -\frac{\partial F}{\partial y}, \quad \frac{dF}{dz} = \frac{\partial F}{\partial z}. \tag{3.21}$$

Hamilton's principal function for the z evolution, i.e., the integral of $p_y\,dy - F\,dz$ along an orbit, is just the usual reduced action,

$$R(y'',z'';y',z') = S(x'',x',E). \tag{3.22}$$

The transformation $x \to (y,z)$ need not be orthogonal, nor even linear, as we can see if we examine the transformation properties of the two sides of Eq. (3.14). For the action S transforms as a scalar under point transformations, and the amplitude determinant D as a pseudoscalar, i.e., like a density, in both the variables x' and x''. Therefore, if we agree to transform G as a half-density in both x' and x'', i.e., as an object which transforms as the square root of a density, then both sides of Eq. (3.14) are seen to be covariant under arbitrary point transformations. But this is exactly how $G(x'',x',E)$ should transform, given that it is the matrix element $\langle x''|(E-H)^{-1}|x'\rangle$. Therefore, we can rewrite Eqs. (3.14) and (3.15) in terms of arbitrary configuration space coordinates, simply by replacing x' and x'' everywhere by the new coordinates.

With these understandings about the general nature of the coordinates (y,z), we may now proceed to derive Eq. (3.20). The proof is only a slight modification of Gutzwiller's original derivation, so we will be sketchy. That is, we write $p' = -\partial S(x'',x',E)/\partial x'$ and $p'' = +\partial S(x'',x',E)/\partial x''$, we make the replacements $x' \to (y',z')$ and $x'' \to (y'',z'')$, and we differentiate $H(x',p') = E$ with respect to y'',z'',E, and $H(x'',p'') = E$ with respect to y',z',E, thereby obtaining

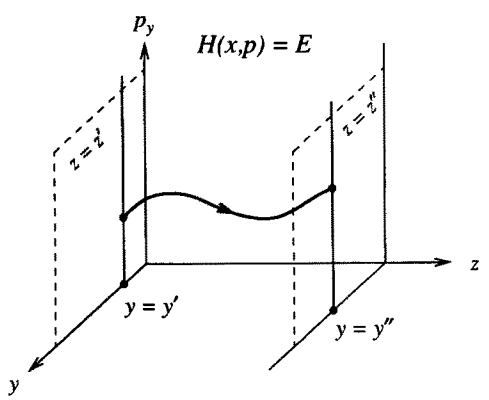

FIG. 6. The Green's function $G(x'',x',E)$ is closely related to a "propagator" $K(y'',z'';y',z';E)$ in the energy shell, in which z is the variable of evolution and $x = (y,z)$. The semiclassical expression for the Green's function is based on orbits taking one from configuration point $y = y'$ in the initial surface of section $z = z'$ to point $y = y''$ in the final surface of section $z = z''$.

six equations. These can be used to reduce the $(f+1)\times(f+1)$ determinant of Eq. (3.15), producing that of Eq. (3.20).

It is instructive now to gather the pieces together, and to write the semiclassical formula for the Green's function in the form,

$$i\hbar G(y'',z'';y',z';E) = \frac{1}{(2\pi i\hbar)^{(f-1)/2}} \sum_r \frac{1}{\sqrt{|\dot z'\dot z''|}} \left|\det \frac{\partial^2 S}{\partial y''\,\partial y'}\right|^{1/2} \times \exp\left[\frac{i}{\hbar} R(y'',z'';y',z') - i\mu\frac{\pi}{2}\right]. \tag{3.23}$$

The most striking feature of this formula is the fact that the right side, apart from the factor $1/\sqrt{|\dot z''\dot z'|}$, is exactly what one would expect for the "propagator" for the z evolution in the energy shell, taking one from the initial surface of section $z = z'$ to the final one $z = z''$. Of course, in classical mechanics, the surface of section is a phase space in its own right of $(f-1)$ degrees of freedom, and we might therefore think to associate it with a Hilbert space of wave functions $\psi(y)$. These wave functions could be thought of as evolving in the variable z according to a Schrödinger equation involving the Hamiltonian F. Then the semiclassical formula for the propagator for the z evolution would be the Van Vleck formula, Eq. (3.1), with f replaced by $f-1$, x by y, and t by z. The result would be precisely Eq. (3.23), apart from the factor $1/\sqrt{|\dot z'\dot z''|}$.

This picture cannot be taken too literally, however, since in general the surface of section has no quantal analog. That is, only certain phase spaces can be quantized, depending on topological and geometrical features. Such considerations are especially important for compact phase spaces, and surface of sections are often compact. Nevertheless, if a surface of section could be quantized, then the semiclassical formula for the propagator would be the Van Vleck formula, as we have described.

The factor $1/\sqrt{|\dot z'\dot z''|}$ spoils the propagator interpretation, but one might say that this factor is present because the real z evolution is governed by H, not $F = -p_z$. That is, if we take the wave equation for $G(x,x',E)$ to be $(\hat H - E)G = 0$, valid for $x \neq x'$, and we interpret it as an evolution equation in z, then the equation is not first order in the z derivatives; whereas the propagator $K(y,z;y',z')$ for the z evolution, as we have described it, would satisfy $(\hat F - i\hbar\,\partial/\partial z)K = 0$ (again for $x \neq x'$), which is first order in the z derivatives. The interpretation of an equation such as $(\hat H - E)G = 0$ as an evolution equation in z has much in common with the paraxial approximation in optics,[33] which follows essentially the same logic.

Let us now interpret $G(x,x',E)$ within the framework of the general structure of semiclassical matrix elements developed in Sec. II. We continue to work with the ordinary phase space and energy shell. If we avoid the point $x = x'$, then $G(x,x',E)$, regarded as a wave function in x, satisfies the time-independent Schrödinger equation as shown in Eq. (3.13). Therefore, the semiclassical expression for G, Eq. (3.14), must have the formal structure of an energy eigen-

function of energy E, and we expect it to be a special case of Eq. (2.8) or Eq. (2.12).

In this regard, it is interesting that the Green's function is defined for energies E which are not eigenvalues, so that the semiclassical representations of G are supported by Lagrangian manifolds that are not quantized, even when the system possesses bound states. That this is consistent is due to two facts. The first is that the inhomogeneous term in the wave equation for G effectively cuts off negative time orbits by introducing a discontinuity in the Lagrangian manifold; and the second is that, with a positive imaginary part to the energy, the infinite sum over branches of the semiclassical wave function, representing positive time orbits, gives a convergent series. For example, in one degree of freedom, the infinite sums in the semiclassical Green's function have the form

$$\sum_{n=0}^{\infty} e^{in\phi} = \frac{1}{1-e^{i\phi}}, \tag{3.24}$$

where $\phi = S/\hbar - \mu\pi/2$. Here, S is the action of the bound orbit $H = E$ and μ is the usual Maslov index of the orbit; S has a small positive imaginary part if E does, so the series converges. It is not necessary for ϕ to be a multiple of 2π (the usual Bohr–Sommerfeld quantization condition), and, indeed, if ϕ approaches such a value, the series diverges, giving the required pole of the Green's function at the energy eigenvalue. An interesting analysis of such series in one degree of freedom has been given by Miller,[34] who includes tunnelling orbits; this analysis has been extended and applied further by Robbins, Creagh, and Littlejohn.[35]

In any case, the semiclassical Green's function of Eq. (3.14) must be supported by an invariant Lagrangian manifold, i.e., a Lagrangian manifold imbedded in the energy shell $H = E$. Figure 6 correctly suggests that the Lagrangian manifold in question must be the one created by allowing the initial Lagrangian manifold $y = y'$ in the surface of section $z = z'$ to sweep out with the flow into the energy shell. The resulting Lagrangian manifold is illustrated in Fig. 7, which

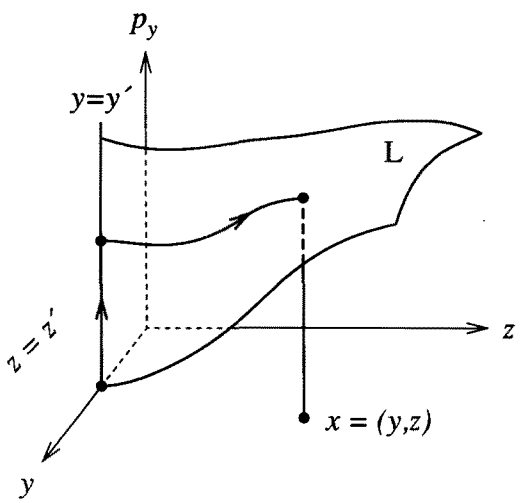

FIG. 7. The Lagrangian manifold in the energy shell $H = E$ supporting the Green's function $G(x,x',E)$ is obtained by mapping the vertical line $y = y'$ forward from the surface of section $z = z'$. Its generating function is the reduced action $S(x,x',E)$.

is almost the same as Fig. 5, apart from changes in symbols and interpretation. One difference, however, is that for systems with bound states, especially when the classical motion is chaotic, the Lagrangian manifold of Fig. 7 will develop complicated self-intersections. These do not occur in Fig. 5, because time always flows forward. The generating function of the Lagrangian manifold in Fig. 7 is simply the reduced action $S(x,x',E)$, as is easily seen by integrating along contours, much as was done in Fig. 5. The Lagrangian manifold of Fig. 7 develops a caustic of order $f-1$ as $x \to x'$; this is not as severe as the caustic of the propagator at $t = t'$. In one degree of freedom, the Green's function has no caustic at all at short distances. Asymptotic forms for the short distance Green's function are best obtained by going to a momentum representation, in accordance with Maslov's philosophy; it is not necessary to use Hankel functions, as many authors have done.

An obvious set of f commuting constants of motion (the A's, of Sec. II) to associate with the Lagrangian manifold of Fig. 7 is the set (Y',H), where Y' represents the $f-1$ functions of (x,p) obtained by following the orbit passing through (x,p) backwards until it crosses the surface of section $z = z'$, and noting the value of $y = y'$ there. The functions Y' are constants of motion because they are constant along orbits, but they are either multivalued or discontinuous constants of motion, of the type discussed below Eq. (2.8). We will denote the values of (Y',H) by (y',E), which we identify with the a's of Sec. II. In this interpretation, the quantity z', like the initial time t' in Eq. (3.10), is merely a parameter.

If we now apply Miller's formula, Eq. (2.8), to the matrix element $\langle y,z|y',E\rangle$, we find that it does not quite reproduce Eq. (3.14) for the Green's function, due to the amplitude prefactor. That is, Eq. (2.8) produces the determinant,

$$\det\begin{pmatrix} \dfrac{\partial^2 S}{\partial y\,\partial y'} & \dfrac{\partial^2 S}{\partial y\,\partial E} \\ \dfrac{\partial^2 S}{\partial z\,\partial y'} & -\dfrac{\partial^2 S}{\partial z\,\partial E} \end{pmatrix} = \frac{1}{\dot z}\det\frac{\partial^2 S}{\partial y\,\partial y'}, \tag{3.25}$$

as we show in the manner of the derivation of Eq. (3.20). Thus the amplitude for $\langle y,z|y',E\rangle$ differs from that in the Green's function by the factor $1/\sqrt{\dot z'}$. But this factor is a constant along orbits, i.e., it can be interpreted as a constant of motion and identified with $F(x,p)$ in Eq. (2.13). Thus we see that the semiclassical expression for $G(x,x',E)$ cannot be represented, even formally, as a unitary matrix element of the form $\langle x|a\rangle$. Instead, it is necessary to use the more general form for semiclassical eigenfunctions given in Eq. (2.13).

A final comment about Gutzwiller's formula, Eq. (3.14), concerns the generality of the coordinates used in it. Above it was shown that this formula is covariant under arbitrary point transformations, but in fact it must be covariant under arbitrary canonical transformations. The justification for this assertion is the general covariance of semiclassical mechanics under canonical transformations, as shown originally by Miller.[10] For example, Eq. (3.14) for the semiclassical Green's function can easily be rewritten in coordinates tied to the stable and unstable manifolds of an unstable

periodic orbit. Such coordinates simplify the amplitude determinant and therefore some of the steps one must follow in taking the trace to get the density of states. However, these coordinates also have certain drawbacks, such as the fact that they are discontinuous in phase space, so we will not pursue this idea further here.

IV. SEMICLASSICAL STRUCTURE OF TRACE FORMULAS

The algebraic manipulations involved in the stationary phase approximation follow a few well-defined patterns, which are essentially the same from one application to the next. The geometrical structures associated with these operations, however, may be quite different. This is immediately apparent when we contemplate Gutzwiller's stationary phase calculation of the trace of the Green's function, and try to understand it in terms of the theory discussed in Sec. II, in which stationary phase points are associated with intersections of Lagrangian manifolds. The immediate problem is that it is not evident how two Lagrangian manifolds are associated with a single Green's function, or how periodic orbits are to be identified with the intersection of anything with anything else.

To clarify the role of the stationary phase approximation in the process of taking traces, we will work with the vector space of operators, in much the same manner as we have previously worked with the vector space of Schrödinger wave functions. We may regard this "operator space" as a kind of doubled Hilbert space; more precisely, for a suitable class of operators it is the tensor product of the Hilbert space of Schrödinger wave functions with its dual. For example, if a basis $|n\rangle$ of Schrödinger wave functions is chosen, then we have an associated basis $|n\rangle\langle m|$ of operators.

For a given operator A, we will talk about the associated "doubled wave function" or "wave function of the operator A," defined by

$$\psi_A(x,x') = \langle x|A|x'\rangle. \tag{4.1}$$

This wave function will be regarded as living on a doubled configuration space, with coordinates (x,x'). By change of basis, it can also be regarded as living on other doubled spaces, such as (x,p'), (p,x'), or (p,p'). In terms of these doubled wave functions, the Hilbert–Schmidt scalar product of operators can be written

$$\mathrm{Tr}(A^{\dagger}B) = \int dx\,dx'\,\psi_A(x,x')^*\psi_B(x,x'). \tag{4.2}$$

In this manner, traces of operators appear as scalar products of wave functions, and can be incorporated into the formalism of Sec. II.

We will also be interested in the semiclassical expressions for the doubled wave functions of various operators. When these expressions exist, they are of the form

$$\psi(x,x') = \sum_r \Omega(x,x')\exp\left[\frac{i}{\hbar}S(x,x') - i\mu\frac{\pi}{2}\right], \tag{4.3}$$

for some amplitude $\Omega(x,x')$ and action $S(x,x')$. Such a semiclassical expression is interpreted geometrically in terms of a Lagrangian manifold in a doubled phase space of $4f$ dimensions, whose coordinates are (x,p,x',p'). For example, the stationary phase points of the integral in Eq. (4.2) can be viewed as intersections of two Lagrangian manifolds in the doubled phase space. As we will show below, Gutzwiller's periodic orbits are intersections of this kind.

We begin this section by examining the doubled wave functions of unitary operators and their semiclassical representation in the doubled phase space. Unitary operators not only possess the simplest doubled wave functions, but also provide us with everything we need to study a variety of trace formulas. It turns out that Lagrangian manifolds in the doubled phase space play a dual role; not only do they support doubled wave functions, but they are also graphs of canonical transformations. The interplay between these roles runs throughout the asymptotics of doubled wave functions, and provides a richness that has no analog in the case of ordinary (single) wave functions and phase spaces. We finally use the geometrical structures in the doubled phase space to reduce a variety of results, including a trace formula due to Tabor,[3] an intermediate result of Berry's[4] on the scars of Wigner functions in phase space, and the Gutzwiller–Balian–Bloch trace formula, to special cases of the asymptotic scalar products presented in Eqs. (2.16) and (2.27).

A. Unitary operators and the doubled phase space

We will primarily be concerned with unitary operators and their doubled wave functions. When these wave functions have a semiclassical limit, it is given by standard WKB theory, most conveniently expressed in terms of Miller's formula for the scalar product $\langle a|b\rangle$. A matrix element $\langle x|U|x'\rangle$ of a unitary operator can always be written as a scalar product of the form $\langle a|b\rangle$ by interpreting the operator U in a passive sense (i.e., as representing a change of basis) rather than an active sense (i.e., as representing a unitary mapping from one basis to another). This is precisely what we did for the propagator in deriving Eq. (3.4). Therefore, unitary operators provide us with examples of doubled wave functions, in which we can read off the amplitude $\Omega(x,x')$ and action $S(x,x')$. We will now consider three specific examples.

The first is the propagator $K(x,t;x',t')$, in which the times t,t' are considered parameters and are required to satisfy $t > t'$. The propagator is the doubled wave function of the unitary time evolution operator $U(t,t')$, so we may write $\psi_U(x,x')$ for it. Furthermore, we have a ready-made semiclassical expression for this wave function, namely, the Van Vleck formula, Eq. (3.1). From this we can easily read off the amplitude $\Omega(x,x')$ and action $S(x,x')$. Notice that in our previous discussion of the Van Vleck formula we regarded x as the variable upon which the wave function depended, and treated x' as a parameter, whereas now we are treating x and x' on an equal footing. This seems like a small change, but it has interesting consequences, as we shall see.

A second example is simply the identity I, which has the doubled wave function

$$\psi_I(x,p') = \frac{e^{ixp'/\hbar}}{(2\pi\hbar)^{f/2}}. \tag{4.4}$$

By writing this in the (x,p') representation, we see that the semiclassical expression is exact, and we can again read off the amplitude and action. Had we used the (x,x') representation, we would have had a caustic, i.e., the wave function would be a delta function concentrated on the line $x = x'$ in the doubled configuration space.

A final unitary operator of interest is $W(\bar{x},\bar{p})$, where $(\bar{x},\bar{p})$ are parameters. It is defined by

$$\psi_W(x,p') = \langle x|W(\bar{x},\bar{p})|p'\rangle = \frac{1}{(2\pi\hbar)^{f/2}} \exp\left\{\frac{i}{\hbar}[2\bar{p}(x-\bar{x}) + p'(2\bar{x}-x)]\right\}. \quad (4.5)$$

The doubled wave function shown is essentially the kernel of the Weyl transform, for if $\hat{A}$ is an operator and $A(\bar{x},\bar{p})$ the corresponding Weyl symbol,[36] then

$$A(\bar{x},\bar{p}) = 2^f \mathrm{Tr}[W(\bar{x},\bar{p})^\dagger \hat{A}]. \quad (4.6)$$

Again, in order to avoid caustics, we choose the (x,p') representation in writing Eq. (4.5), and again, the semiclassical approximation is exact.

All three of these unitary operators, $U(t,t')$, I, and $W(\bar{x},\bar{p})$, correspond in a simple way to classical canonical transformations. As shown by Miller,[10] this correspondence is realized through the actions $S(x,x')$ or $S(x,p')$, treated either as a F_1- or F_2-type generating function,[37] respectively, for a canonical transformation connecting (x,p) with (x',p'). One slight subtlety is that (x',p'), which appear as initial variables in the propagator, must be treated as "new" variables, whereas the final (x,p) are "old" variables.

For example, the action of the Van Vleck formula is Hamilton's principal function, the F_1-type generating function taking us from final (x,p) to initial (x',p'); and the action of the doubled wave function for the identity operator, Eq. (4.4), is $F_2(x,p') = xp'$, the generator of the classical identity canonical transformation. As for the unitary operator $W(\bar{x},\bar{p})$, the generating function is

$$S(x,p') = F_2(x,p') = 2\bar{p}(x-\bar{x}) + p'(2\bar{x}-x), \quad (4.7)$$

which generates the "averaging" canonical transformation, given implicitly by

$$\bar{x} = (x+x')/2, \quad \bar{p} = (p+p')/2. \quad (4.8)$$

Let us now consider the doubled phase space as a medium for interpreting the semiclassical expressions for our doubled wave functions. There is a certain subtlety concerning this space that requires some elaboration.

To begin, let us take a semiclassical representation for a doubled wave function, as in Eq. (4.3), and let us plot the wave fronts or contours of $S(x,x')$ in the doubled configuration space. This is illustrated schematically in Fig. 8. The gradient of $S(x,x')$ in the doubled configuration space is a vector perpendicular to the wave fronts; let us provisionally write $p = \partial S/\partial x, p' = \partial S/\partial x'$ for its components. These are functions of (x,x'), and give us a momentum field on the configuration space. Then in accordance with concepts from the WKB theory that encompass all kinds of wave equations, not just those of quantum mechanics, we construct a phase

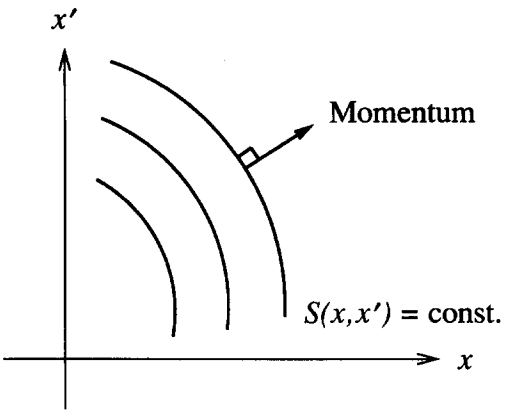

FIG. 8. The momentum can be defined as the gradient of the action, in this case in the doubled configuration space. If, however, p and p' are defined, respectively, as the old and new momenta under the canonical transformation $(x,p) \to (x',p')$, then the components of this momentum vector are $(p, -p')$.

space with coordinates (x,p,x',p'), within which our specific momentum field for our specific wave function is represented by a Lagrangian manifold. In this phase space, the variables conjugate to (x,x') are (p,p'), respectively, as defined above; it is in terms of these variables that Hamilton's equations for the rays take on their standard form and that Poisson brackets are computed in the standard way.

On the other hand, we are also identifying the action $S(x,x')$ with the F_1-type generating function of a canonical transformation, for which the generating function relations are $p = \partial S/\partial x, p' = -\partial S/\partial x'$. The point of this is that the equation for p' has a minus sign relative to the definition of p' in the preceding paragraph. The reason for this discrepancy is that there are now two conflicting interpretations for the symbol p': it is either the momentum conjugate to x' in the doubled phase space, or else it is the new momentum in the canonical transformation. We can choose only one of these interpretations, so by convention we will take the latter: our p' will be the new momentum in the canonical transformation. (We abandon the definition of p' in the preceding paragraph.)

This means that in the doubled phase space, the momentum conjugate to x' is $-p'$, and that the vector perpendicular to the wave fronts in the doubled configuration space has components $(p, -p')$. Therefore, the action differential on the doubled phase space is

$$dS = p\,dx - p'\,dx', \quad (4.9)$$

and the symplectic form is

$$\omega_D = dp \wedge dx - dp' \wedge dx'. \quad (4.10)$$

A Lagrangian manifold in the $4f$-dimensional doubled phase space is a $2f$-dimensional manifold on which this form vanishes. Finally, if $f(x,p,x',p')$ and $g(x,p,x',p')$ are any two functions on the doubled phase space, then their Poisson bracket is given by

$$\{f,g\}_D = \left(\frac{\partial f}{\partial x}\frac{\partial g}{\partial p} - \frac{\partial f}{\partial p}\frac{\partial g}{\partial x}\right) - \left(\frac{\partial f}{\partial x'}\frac{\partial g}{\partial p'} - \frac{\partial f}{\partial p'}\frac{\partial g}{\partial x'}\right) = \{f,g\} - \{f,g\}', \quad (4.11)$$

where the D subscript indicates the doubled Poisson bracket, where the primed Poisson bracket is taken only with respect to the variables (x',p'), and where the unprimed Poisson bracket is taken only with respect to (x,p).

We now effectively have three phase spaces, an unprimed, a primed, and a doubled, which we may denote by Φ, Φ', and Φ_D, respectively. We may also denote the creation of Φ_D out of the other two, in accordance with Eq. (4.11), by $\Phi_D = \Phi \otimes \Phi'^*$. The reason for this notation is analogy: if we denote the Hilbert space of Schrödinger wave functions by X, then the Hilbert space of operators is $X \otimes X^*$. We may also note that the minus sign in the nonstandard Poisson bracket of Eq. (4.11) comes originally from a complex conjugation of a phase.

The doubled phase space with the symplectic structure of Eq. (4.10) is a well-known device for representing the geometrical structure of canonical transformations and their generating functions. It is discussed, for example, by Abraham and Marsden.[11] It is possible to give a geometrical representation of a canonical transformation without going to the doubled phase space; for example, a foliation of the $2f$-dimensional single phase space into an f-parameter family of Lagrangian manifolds does this. But such a representation does not treat the old and new variables on an equal footing, nor does it readily reveal the relationship among the various kinds of generating functions for a given canonical transformation. To satisfy these goals, the doubled phase space is necessary.

The doubled phase space is used to represent canonical transformations in the following manner. Suppose we are given a canonical transformation, say,

$$x = X(x',p'), \quad p = P(x',p'). \tag{4.12}$$

Then the set of points (x,p,x',p') in the doubled phase space that satisfy these equations, i.e., the graph of the canonical transformation, is a $2f$-dimensional surface in this space. The same surface can be equally well specified by the various generating function relations; for example, if the canonical transformation possesses an F_1-type generating function, then the equations

$$p = \frac{\partial F_1(x,x')}{\partial x}, \quad p' = -\frac{\partial F_1(x,x')}{\partial x'} \tag{4.13}$$

are equivalent to Eqs. (4.12), being simply an algebraic rearrangement of them.

It turns out that this $2f$-dimensional surface is a Lagrangian manifold in the doubled phase space with respect to the symplectic form of Eq. (4.10). This is most easily seen with the aid of a diagram such as Fig. 9. We let L be the graph of the given canonical transformation, and we choose a contractible closed curve in L, bounding a two-dimensional region Γ_D. This region is projected onto the (x,p) and (x',p') phase spaces, producing, respectively, regions Γ and Γ'. Since L is the graph of the canonical transformation, the region Γ' is the image of the region Γ under the canonical transformation. But since this transformation is canonical, the respective symplectic areas A, A' of Γ, Γ', measured in the unprimed and primed phase spaces, must be equal. This, in turn, implies

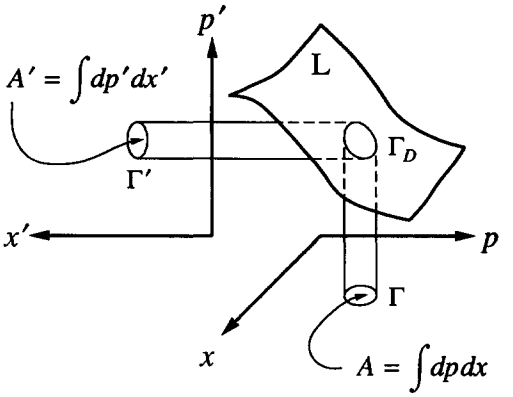

FIG. 9. The graph of a canonical transformation $(x,p) \to (x',p')$ in the doubled phase space is a Lagrangian manifold L with respect to the symplectic form of Eq. (4.10). Conversely, every Lagrangian manifold in this space with a nonsingular projection onto the (x,p) and (x',p') planes is the graph of a canonical transformation.

$$\int_{\Gamma_D} \omega_D = A - A' = 0. \tag{4.14}$$

Since this is true for arbitrary contractible Γ_D on L, we must have $\omega_D = 0$ on L, and therefore L is Lagrangian with respect to ω_D.

Conversely, suppose we are given a Lagrangian manifold L in the doubled phase space. A region of L that has a nonsingular projection onto the (x,p) and (x',p') phase spaces provides a mapping between these phase spaces; and this mapping preserves the symplectic area of two-dimensional area elements. Therefore, the mapping is canonical, at least within the given regions. [One can show that a region of a Lagrangian manifold in the doubled space has a nonsingular projection onto (x,p) if and only if its projection onto (x',p') is also nonsingular.]

A given Lagrangian manifold in the doubled phase space may well possess points where the projections onto the (x,p) and (x',p') planes are singular. In the typical case, these singularities occur on a subset of L of lower dimensionality than $2f$, and represent the places where the branches of a multivalued canonical transformation coalesce. We may note, however, that the three canonical transformations discussed above, the time evolution, the identity, and the averaging transformation, are all single valued, so the corresponding Lagrangian manifolds have nonsingular projections onto (x,p) and (x',p'). It is also possible to construct exceptional Lagrangian manifolds in the doubled phase space which have projections onto (x,p) and (x',p') which are singular everywhere, and which therefore do not represent canonical transformations.

The representation of canonical transformations by Lagrangian manifolds in the doubled phase space is useful for insight it offers into the various types of generating functions. For example, if the Lagrangian manifold corresponding to a given canonical transformation has a nonsingular projection onto the (x,x') plane, then it has an F_1-type generating function. In this case, the generating function of the canonical transformation coincides with the generating function of an isolated Lagrangian manifold, as was dis-

cussed below Eq. (2.2), if the latter concept is transplanted to the doubled phase space. Notice that the projection onto (x,x') may be multivalued and have singularities, even if the projection onto (x,p) and (x',p') does not. This circumstance occurs, for example, in the canonical transformation for the time evolution.

More generally, we can obtain a generating function of any type by projecting L onto some $2f$-dimensional plane, spanned by some subset of old and new x's and p's. The only requirement is that the projection must be nonsingular, and that the plane onto which we project must itself be Lagrangian with respect to the doubled symplectic form. For example, (x,p') is allowed, but (x,p) is not.

One of the most important applications of Lagrangian manifolds in physical problems is in the representation of multidimensional wave fields in the short wavelength limit. The preceding discussion of the doubled phase space has concerned another application, namely, an elegant and symmetrical means of treating canonical transformations and all of their generating functions. It is remarkable, therefore, that these two applications come together in the asymptotics of the matrix elements of operators, i.e., in our doubled wave functions. We see now that every matrix element with a semiclassical limit corresponds to a Lagrangian manifold in the doubled phase space, which in turn, if it has nonsingular projections onto the (x,p) and (x',p') subspaces, corresponds to a canonical transformation.

Although in this paper we are primarily interested in doubled wave functions of unitary operators, other operators may also be considered, and they are also represented semiclassically by Lagrangian manifolds in the doubled phase space. It is interesting that some of these give rise to semiclassical wave functions that do not correspond to a canonical transformation, because their Lagrangian manifolds have projections onto (x,p) or (x',p') space which are singular everywhere. For example, let $\psi(x) = \langle x|\psi\rangle$ be an ordinary (single) wave function with a semiclassical representation in the usual form, and consider the projection operator $|\psi\rangle\langle\psi|$. The doubled wave function for this projection operator gives rise to the Lagrangian manifold,

$$p = f(x), \quad p' = f(x') \tag{4.15}$$

in the doubled phase space, where $f(x) = \partial S(x)/\partial x$, $S(x)$ being the (single) action of $\psi(x)$. This Lagrangian manifold has a singular projection onto (x,p) space, because only certain (x,p) values are consistent with Eq. (4.15), and because once such an (x,p) value is given, the values of (x',p') are undetermined. The Lagrangian manifold of Eq. (4.15) does have an F_1-type generating function; however, it is $F_1(x,x') = S(x) - S(x')$, and it is an example of a generating function that does not generate a canonical transformation. Obviously such generating functions must be taken into account in a general theory of semiclassical matrix elements.

B. The Hilbert–Schmidt scalar product of unitary operators

We will now apply the results of Sec. II, especially Eq. (2.16) for the matrix element $\langle a|b\rangle$, to the computation of $\mathrm{Tr}(V^\dagger U)$, in which U and V are unitary operators. In order to have a specific example in mind, we may identify V with the identity operator I, and U with $U(t,t')$, the unitary time-evolution operator for a time-dependent system, so as to compute $\mathrm{Tr}\, U(t,t')$. We will assume that the system has no constants of motion, either the Hamiltonian or anything else. The reason for this assumption is that it leads to the simplest trace formula providing a connection with classical periodic orbits, and therefore is a good place to start. In addition, we have seen in Sec. III that the Green's function is very nearly a propagator in the surface of section, so taking traces of propagators is good practice for taking traces of Green's functions.

Actually, in computing $\mathrm{Tr}(V^\dagger U)$, it is sufficient to assume $V = I$, since we can always rewrite the operator product $V^\dagger U$, itself unitary, simply as U. This step simplifies the calculation somewhat, and, when we are done, we are always free to factorize U once again into the product of two unitary operators. Let us therefore proceed with this simplification.

In the semiclassical computation of $\mathrm{Tr}(I^\dagger U)$, we know that the asymptotic forms for the matrix elements of I and U are associated with two Lagrangian manifolds in the doubled phase space, which we denote by L_I and L_U; and that these in turn are associated with two canonical transformations, the I transformation, which is the identity, and the U transformation, which we denote by $x = X(x',p')$, $p = P(x',p')$. We will denote the F_1-type generating function of the U transformation by $S(x,x')$, which is also the action for the semiclassical expression for $\psi_U(x,x') = \langle x|U|x'\rangle$. The amplitude of this expression is given by Miller's formula, Eq. (2.8), as

$$\Omega(x,x') = \left|\det \frac{\partial^2 S(x,x')}{\partial x\, \partial x'}\right|^{1/2}. \tag{4.16}$$

An immediate fact that emerges from a consideration of Eq. (2.16) is that the stationary phase points in the computation of $\mathrm{Tr}(I^\dagger U)$ will be the intersections of L_I and L_U in the doubled phase space, i.e., the points (x,p,x',p') such that

$$x = x' = X(x',p'), \quad p = p' = P(x',p'). \tag{4.17}$$

These points on the doubled space represent periodic orbits on the single space, in the sense that $(x,p) = (x',p')$ is a fixed point of the U-canonical transformation. Another immediate fact is that the action at one of these stationary phase points is simply $S(x,x')$, because the (x,x') action of the identity transformation is zero. [Properly, one should work in a representation in which neither operator has a caustic; it does not matter which, just as in Eq. (2.16) the right-hand side is obviously independent of the choice of the x representation. However, the results are as quoted here.]

To proceed further with the application of Eq. (2.16), however, we must somehow represent the doubled wave function $\psi_U(x,x')$ as the simultaneous eigenfunction of some set of commuting operators, say, $\mathscr{B}$, with eigenvalues b and classical counterparts B; and similarly the doubled wave function for the identity, using operators $\mathscr{A}$, eigenvalues a, and classical counterparts A. Because we are now working in the doubled space, each one of these sets contains $2f$ members, and the operators $\mathscr{A}$, $\mathscr{B}$ are not ordinary operators, but rather "doubled operators," i.e., linear mappings of ordinary operators into other ordinary operators. Similarly,

the sets of classical functions A, B are functions of (x,p,x',p'), and must Poisson commute within each set according to the doubled Poisson bracket.

We must do this because on the doubled phase space, the doubled wave functions are represented by two particular Lagrangian manifolds in isolation, which are not members of foliations. Therefore, in order to apply Eq. (2.16), or rather its transcription to the doubled phase space, it is necessary to imbed these individual Lagrangian manifolds in foliations. The analog of this process in the case of the single phase space would be to go from a particular solution of the Hamilton–Jacobi equation, say $S(x)$, to a complete solution, say $\tilde{S}(x,a)$, for which $S(x) = \tilde{S}(x,a)$ for some fixed value of a ($a = 0$ is convenient). On the doubled phase space, we may regard the particular Lagrangian manifolds L_I and L_U as the "physical" Lagrangian manifolds, and the others as being used simply for the sake of applying Eq. (2.16). Since (virtually) all Lagrangian manifolds in the doubled space correspond to canonical transformations, each of the two foliations will produce two families of canonical transformations, parametrized by a or b.

The proposed imbedding is not unique and seems artificial, since a straightforward application of the stationary phase approximation, along the lines of Gutzwiller's original derivation, does not require it, but rather works (effectively) with the two physical Lagrangian manifolds and the densities on them. We will proceed anyway with the suggested approach for the following reasons. First, the manifest phase space invariance of Eq. (2.16) is compelling, and reason to pursue this formula to see how general it is. Second, when we imbed a particular Lagrangian manifold in a foliation and use the foliation in the manner suggested, it turns out that only the members of the foliation in an infinitesimal neighborhood of the original Lagrangian manifold have an effect on the result (as one would expect). That is, these neighboring Lagrangian manifolds and the manner in which they are specified are equivalent to positing a density on the original Lagrangian manifold; therefore, working with such a foliation in the neighborhood of the original Lagrangian manifold is an alternative to dealing with densities and amplitude determinants. Finding a means of clarifying manipulations on amplitude determinants is a major goal of this paper. Third, the foliation we will suggest below leads to an interesting and novel means of calculating Gutzwiller's periodic orbit amplitudes. Fourth, the Wigner–Weyl formalism provides us for free with a foliation of the desired kind, and the results of pursuing this line of investigation yield new insights into the Wigner–Weyl formalism (and into Berry's[4] trace formula for the scars of Wigner functions in phase space). And finally, the kind of thing we propose here is not unheard of in physics; for example, Dirac brackets are usually computed using coordinates defined on a space larger than the physical one.

Let us work with the manifold L_U, and find the desired imbedding; that for L_I will then follow easily. We require $2f$ functions $B_1(x,p,x',p'), \ldots, B_{2f}(x,p,x',p')$ that take on constant values on L_U, and which commute with one another under the doubled bracket. In addition, they should give the correct amplitude determinant $\Omega(x,x')$, shown in Eq. (4.16), in accordance with the doubled version of Miller's formula, Eq. (2.8). We satisfy the first two goals by applying some guesswork, and find

$$B_i(x,p,x',p') = x_i - X_i(x',p'),$$
$$B_{i+f}(x,p,x',p') = p_i - P_i(x',p'), \tag{4.18}$$

for $i = 1, \ldots, f$. These B's take on constant values on L_U, namely $b = 0$; and their doubled Poisson brackets among themselves vanish, as verified by direct substitution into Eq. (4.11).

Consider now members of the foliation for which $b = (b_x,b_p) \neq 0$. The Lagrangian manifold given by $B = b$ represents a canonical transformation, given in terms of the one for $b = 0$ by

$$x = X(x',p') + b_x, \quad p = P(x',p') + b_p, \tag{4.19}$$

which can be regarded as the composition of the U-canonical transformation with a rigid displacement in phase space. This canonical transformation has the F_1-type generating function,

$$\tilde{S}(x,x';b) = S(x - b_x,x') + b_p x, \tag{4.20}$$

as follows from the fact that the F_1-type generating function of the composition of two canonical transformations is just the sum of the F_1-type generating functions of the constituents. The tilde distinguishes the generating function parametrized by b from the the original one (we have $\tilde{S} = S$ when $b = 0$). Therefore, if we apply the doubled version of Miller's formula, Eq. (2.8), to the computation of the simultaneous eigenfunction of the $2f$ doubled operators $\mathscr{B}$ corresponding to the classical B's of Eq. (4.18), we find

$$\langle\langle x,x'|b\rangle\rangle = \text{const} \sum_r \tilde{\Omega}(x,x';b) \times \exp\left[\frac{i}{\hbar}\tilde{S}(x,x';b) - i\mu\frac{\pi}{2}\right], \tag{4.21}$$

where we use double angle brackets for a double scalar product, and where

$$\tilde{\Omega}(x,x';b)^2 = \det\begin{pmatrix} \dfrac{\partial^2\tilde{S}}{\partial x\,\partial b_x} & \dfrac{\partial^2\tilde{S}}{\partial x\,\partial b_p} \\ \dfrac{\partial^2\tilde{S}}{\partial x'\,\partial b_x} & \dfrac{\partial^2\tilde{S}}{\partial x'\,\partial b_p} \end{pmatrix} = \det\begin{pmatrix} \dfrac{\partial^2 S}{\partial x^2} & I \\ -\dfrac{\partial^2 S}{\partial x\,\partial x'} & 0 \end{pmatrix} = \det\frac{\partial^2 S}{\partial x\,\partial x'}. \tag{4.22}$$

Here, the tilde on $\tilde{\Omega}$ has the same meaning as on $\tilde{S}$; and when $b = 0$, we see that $\tilde{\Omega}$ agrees with the amplitude Ω of Eq. (4.16). Therefore, we now have

$$\psi_U(x,x') = \langle x|U|x'\rangle = \text{const}\langle\langle x,x'|b = 0\rangle\rangle; \tag{4.23}$$

apart from normalization, which we deal with momentarily, we have completed the classical and semiclassical aspects of imbedding our wave function, Lagrangian manifold, and canonical transformation in the required families.

Let us now consider the doubled operators $\mathscr{B}$ corresponding to the classical B's of Eq. (4.18). First we introduce certain doubled operators associated with ordinary operators. For example, if F is an ordinary operator, we can associate it with a doubled operator $\mathscr{L}_F$ by left multiplication, i.e.,

$$\mathscr{L}_F G = FG, \tag{4.24}$$

where G is any ordinary operator. Similarly, we can associate F with another doubled operator $\mathscr{R}_F$ by right multiplication, i.e.,

$$\mathscr{R}_F G = GF. \tag{4.25}$$

Then we can represent our desired $\mathscr{B}$'s in terms of left and right multiplication by

$$\begin{aligned} \mathscr{B}_i &= \mathscr{L}_{\hat{x}_i} - \mathscr{R}_{U^\dagger \hat{x}_i U}, \\ \mathscr{B}_{i+f} &= \mathscr{L}_{\hat{p}_i} - \mathscr{R}_{U^\dagger \hat{p}_i U}, \end{aligned} \tag{4.26}$$

for $i = 1, \ldots, f$. These $\mathscr{B}$'s commute with one another, as is easily verified, and their simultaneous eigenoperator $U(b)$ is given by

$$\langle x|U(b)|x'\rangle = e^{ixb_p/\hbar}\langle x - b_x|U|x'\rangle, \tag{4.27}$$

where $U = U(0)$. That is, we have

$$\mathscr{B}U(b) = bU(b). \tag{4.28}$$

We note that $U(b)$ is unitary for all values of b. Equation (4.27) immediately allows us to write down the semiclassical expression for $\langle x|U(b)|x'\rangle$ in terms of that for $\langle x|U|x'\rangle$, and it is precisely Eq. (4.21). Therefore, we have now imbedded our original operator U in a family $U(b)$, whose semiclassical and classical representatives are those given above, and identified the members of the family as simultaneous eigenoperators of a complete set of commuting doubled operators.

There remains only the normalization. It is convenient not to use the precise transcription of Eq. (2.7) to the doubled space, but rather to demand that

$$\langle\langle b|b'\rangle\rangle = (2\pi\hbar)^f \delta(b - b'). \tag{4.29}$$

This has the advantage that the normalized, doubled wave function $\langle\langle x,x'|b\rangle\rangle$ is exactly the unitary matrix element $\langle x|U(b)|x'\rangle$, so that

$$\mathrm{Tr}[U(b)^\dagger U(b')] = (2\pi\hbar)^f \delta(b - b'). \tag{4.30}$$

Thus the constants in Eqs. (4.21) and (4.23) are unity.

Repeating the calculation leading from Eq. (4.18) to (4.28) with U replaced by I, B by A, etc., we derive parallel results for the identity canonical transformation. The most important of these is

$$\begin{aligned} A_i(x,p,x',p') &= x_i - x_i', \\ A_{i+f}(x,p,x',p') &= p_i - p_i', \end{aligned} \tag{4.31}$$

for $i = 1, \ldots, f$.

There now arises an interesting point. As discussed earlier, the use of a complete set of commuting observables such as the A's of Eq. (4.31) or the B's of Eq. (4.18), is an alternative to working with a density on a Lagrangian manifold. Nevertheless, it is of interest to actually compute the density on a Lagrangian manifold in the doubled phase space corresponding to a unitary operator. Any number of coordinate systems are available for expressing the density, such as (x,x'), (x,p'), etc., and it is easy to see that the density function with respect to one of these coordinate systems is just the corresponding amplitude or Van Vleck determinant. These density functions can be thought of as arising from projecting a density intrinsic to the Lagrangian manifold down onto some Lagrangian plane in the doubled phase space. There is, however, no reason why the density must be projected onto a Lagrangian plane; for example, it could be projected onto the (x,p) plane. In the single phase space, there would be no point in projecting onto a plane that was not Lagrangian, because no corresponding representation would exist for the wave functions. But in the doubled phase space, it is interesting to do so, because the (x,p) plane has its own intrinsic, invariant measure, namely, the Liouville measure $dp\,dx$. Not surprisingly, we find that the density on a unitary Lagrangian manifold in the doubled phase space, when projected onto the (x,p) plane, is constant, i.e., it agrees with the Liouville measure there. Thus one can regard the doubled wave functions of unitary operators as having the simplest semiclassical structure of any doubled wave functions, in the sense that everything is specified by the Lagrangian manifold alone, since the density is the natural measure provided by the geometry.

It is now easy to compute the semiclassical expression for $\mathrm{Tr}(I^\dagger U)$. Only the amplitude determinant requires any calculation; using Eqs. (4.18) and (4.31), we find

$$\begin{aligned} \det\{A,B\}_D &= \det\begin{pmatrix} -\dfrac{\partial X}{\partial p'} & -\mathbf{I} + \dfrac{\partial X}{\partial x'} \\ \mathbf{I} - \dfrac{\partial P}{\partial p'} & \dfrac{\partial P}{\partial x'} \end{pmatrix} \\ &= \det(\mathbf{M} - \mathbf{I}), \end{aligned} \tag{4.32}$$

where $\mathbf{M}$ is the symplectic matrix $\partial(X,P)/\partial(x',p')$, and where we have taken the transpose after the first equality. This determinant is to be evaluated at the stationary phase points, i.e., the periodic orbits, for which $\mathbf{M}$ becomes the monodromy matrix. Altogether, the result is

$$\mathrm{Tr}\, U = \sum \frac{\exp[(i/\hbar)S(x,x) - i\mu(\pi/2)]}{|\det(\mathbf{M} - \mathbf{I})|^{1/2}}. \tag{4.33}$$

The sum is taken over fixed points (x,p) of the U-canonical transformation.

A trace formula of this type was first derived by Tabor,[3] who quantized the standard map by imbedding it in a continuous time system, and then used the stationary phase approximation to take the trace of $U(0,T)$, where T is the period of the mapping. Our result is slightly more general than Tabor's, in that we work in any number of degrees of freedom, and we make fewer assumptions about the unitary operator U. For example, even if U is a time-evolution operator, the system need not be time periodic, or, if it is, the elapsed time $t - t'$ need not be a period. These constitute minor modifications to Tabor's results, the main point of our presentation being the geometrical structure we reveal.

Equation (4.33) cannot be immediately generalized to the case of time-independent systems, because, for such systems, the monodromy matrix $\mathbf{M}$ has an eigenvector of eigen-

value $+1$, namely, the flow vector along the periodic orbit, $(\dot{x},\dot{p})$. Therefore, the denominator of Eq. (4.33) vanishes in this case, indicating a caustic. The caustic is of the type in which the two Lagrangian manifolds intersect over a region of dimensionality 1, as discussed in Sec. II B and illustrated in Fig. 4. This is the case originally considered by Gutzwiller, and it will be discussed more fully below. For now we simply note that the intersection I of the two Lagrangian manifolds is just the periodic orbit, and that the appropriate scalar product formula is not Eq. (2.16), but rather Eq. (2.27).

Similarly, we must exclude the possibility of constants of motion in Eq. (4.33), because the symmetry groups associated with such constants map orbits into other orbits. In particular, they map periodic orbits into other periodic orbits, thereby creating a continuous family of nonisolated stationary phase points. Evidently, the Hamiltonian, when conserved, has the same effect on the structure of our trace formula as any other constant of motion.

Of course, we may have caustics of Eq. (4.33) even when $\partial H/\partial t \neq 0$ and no constants of motion exist, such as in the case of a periodic orbit of parabolic stability (two eigenvalues of $\mathbf{M}$ equal to $+1$). Typically such stationary phase points are still isolated, and the caustic is of the Airy function (fold catastrophe) type. The formalism based on the Lagrangian manifolds in the doubled phase space makes it clear that this caustic is no different from any other Airy function-type caustic in any other system; facts like this have not been evident in earlier approaches to trace formulas. (See, however, the analysis by Ozorio de Almeida and Hannay,[38] which reveals a surprisingly complex caustic structure in stable periodic orbits whose higher-order iterations are resonant.)

Finally, it is useful to replace U by $V^{\dagger}U$, as suggested earlier, and write out the result. It is

$$\mathrm{Tr}(V^{\dagger}U) = \sum \frac{\exp\{(i/\hbar)[S_U(x,x') - S_V(x,x')] - i\mu(\pi/2)\}}{|\det(\mathbf{M}_U - \mathbf{M}_V)|^{1/2}}. \tag{4.34}$$

The sum is over points (x',p') which, when mapped by the U-canonical transformation to (x,p), and then again by the inverse of the V-canonical transformation, return to (x',p').

C. The Weyl correspondence

Let us now consider the Weyl transform of a unitary operator U, which we denote by $W_U(\bar{x},\bar{p})$. This problem has been considered by Berry[4] in the case that U is the time-evolution operator, as a first step in the derivation of his formula for the scars of Wigner functions in phase space. Our strategy will be to write the desired Weyl transform as the trace of the product of two unitary operators, as in Eq. (4.6), to which we can immediately apply Eq. (4.34), identifying V with the operator $W(\bar{x},\bar{p})$ of Eq. (4.5).

In doing this an interesting point arises, namely, that the parameters $(\bar{x},\bar{p})$ of the kernel of the Weyl transform, i.e., the location in the (single) phase space at which the function W_U is to be evaluated, can be interpreted as a complete set of $2f$ commuting functions of (x,p,x',p'), specifying a foliation of the doubled phase space into Lagrangian manifolds. This follows by using Eq. (4.8) to define $(\bar{x},\bar{p})$ as functions on the doubled phase space, and then by computing the doubled Poisson bracket. Therefore, it is not necessary to imbed the unitary operator $W(\bar{x},\bar{p})$ in a family or its corresponding Lagrangian manifold in a foliation, since that is already done for us. On the other hand, the foliation supplied to us for free by the Weyl formalism is really no different from the one we guessed in Eq. (4.18), for if we use the averaging canonical transformation to define functions A according to

$$A_i(x,p,x',p') = x_i + x_i' = 2\bar{x}_i,$$
$$A_{i+f}(x,p,x',p') = p_i + p_i' = 2\bar{p}_i, \tag{4.35}$$

for $i = 1,\ldots,f$, we see that we have precisely the form we guessed in Eq. (4.18).

The desired Weyl transform is now immediate. The result is

$$W_U(\bar{x},\bar{p}) = 2^f \sum \frac{\exp\{(i/\hbar)[S(x,x') - \bar{p}(x-x')] - i\mu(\pi/2)\}}{|\det(\mathbf{M}+\mathbf{I})|^{1/2}}, \tag{4.36}$$

where $S(x,x')$ is the action of the U-canonical transformation, and where $\mathbf{M} = \mathbf{M}_U$. This is the formula originally derived by Berry.[4] The sum is taken over points (x',p') which, when mapped by the U-canonical transformation, yield points (x,p) satisfying Eq. (4.8), Berry's "midpoint rule." The plus sign in the determinant in the denominator, in contrast to the minus sign in Eq. (4.22), is a simple consequence of replacing the identity canonical transformation by the averaging transformation in the computation of $\det\{A,B\}_D$, giving $\mathbf{M}_V = -\mathbf{I}$.

In Berry's calculation, the derivation of Eq. (4.36) is only the first (and easier) step. The second step consists of taking the Fourier transform in time of Eq. (4.36), to obtain the Weyl transform of the Green's function. Although the second step only involves a one-dimensional integral, it is rather intricate in its execution, leading ultimately to Airy function caustics in the neighborhood of periodic orbits.

In order to limit the scope of this paper, we will not pursue this line of investigation here, but rather promise it in future work. We will, however, offer the following general comments on the relation of the Wigner–Weyl formalism to the doubled phase space and its doubled Poisson bracket, and then make some final comments on Berry's calculation.

The considerations raised here reveal several features of the Wigner function and Weyl correspondence which otherwise are obscure. For example, consider the obvious fact that

the Wigner function can be regarded as a kind of "wave function" on phase space. A peculiar feature of this interpretation, however, is that $(\bar{x},\bar{p})$ behave like configuration space coordinates, not only because they are the variables upon which a wave function depends, but also because they play the role of commuting q's when the asymptotics of the Wigner function are considered. The present analysis shows that $(\bar{x},\bar{p})$ actually are commuting variables, i.e., under the doubled Poisson bracket, and can indeed be taken as a complete set of commuting variables on the doubled phase space. That is, they constitute one-half of a canonical coordinate system on the doubled phase space; the simplest choice for the other half is

$$\tilde{x} = x - x', \quad \tilde{p} = p - p'. \tag{4.37}$$

These give the commutation relations

$$\{\bar{x}_i,\tilde{x}_j\}_D = \{\bar{p}_i,\tilde{p}_j\}_D = 0,$$
$$\{\bar{x}_i,\tilde{p}_j\}_D = \{\tilde{x}_i,\bar{p}_j\}_D = \delta_{ij}, \tag{4.38}$$

showing that $(\bar{x},\bar{p};\tilde{x},\tilde{p})$ are canonical variables on the doubled phase space.

For another example, much is known about the caustic structure of Wigner functions, especially as a consequence of the work of Berry;[39] what the present analysis shows is that these caustics are singularities of the projection of one Lagrangian manifold in the doubled phase space onto another. In other words, the asymptotic structure of the Wigner function can be handled as a special case of Maslov's theory, a fact not completely evident when the stationary phase approximation is applied to the formulas of the Wigner–Weyl formalism. Indeed, Berry[39] regarded the similarities between his calculations and Maslov's theory as more superficial than real.

To elaborate on this point, we may contrast the perspective of this paper on caustics, as properties of the intersections of two Lagrangian manifolds in phase space, with the usual perspective, in terms of the singularities of the projection of one Lagrangian manifold onto another. The two views are equivalent, of course; for example, in a simple case we consider $\psi(x) = \langle x|\psi\rangle$, an ordinary (single) wave function, and regard the caustics as being the singularities of the projection of the Lagrangian manifold L_ψ onto configuration space. The projection takes place along lines of constant x, themselves Lagrangian manifolds whose intersections with L_ψ gives the picture expressed in Fig. 1 and Eq. (2.16). Configuration space can be identified with the subset of phase space $p = 0$, also a Lagrangian plane; but more generally, any surface transverse to the projection would work as well. That is, configuration space may be thought of as the quotient space under the equivalence engendered by the projection.

Similarly, in the doubled phase space, the caustics of the Weyl transform of an operator can be interpreted in terms of the intersections of some Lagrangian manifold [L_U in Eq. (4.36)] with the Lagrangian manifolds corresponding to the averaging transformation; or in terms of the projection of the given Lagrangian manifold onto the plane $\tilde{x} = \tilde{p} = 0$. The projection takes place at constant $(\bar{x},\bar{p})$, which also serve as coordinates on the plane $\tilde{x} = \tilde{p} = 0$. In the usual interpretation of the Weyl transform, this latter plane is identified with the (single) phase space, although it could be equally well identified with the Lagrangian plane specifying the identity canonical transformation, since $\tilde{x} = \tilde{p} = 0$ implies $x = x'$, $p = p'$. One final comment is that the so-called covariant Weyl symbol,[36] obtained by using a Heisenberg operator $T(\bar{x},\bar{p})$ in Eq. (4.6) instead of the operator $W(\bar{x},\bar{p})$, corresponds to projecting the given Lagrangian manifold onto the complementary plane $\bar{x} = \bar{p} = 0$, which is also sometimes identified with the single phase space (or a second copy of it).

These considerations make it evident that the Airy functions and interference fringes found by Berry[4] for the scars of Wigner functions in phase space are a direct consequence of projecting a certain Lagrangian manifold in the doubled phase space onto the Lagrangian plane representing the identity transformation. The proper way to do this is to work in the extended, doubled phase space, with coordinates $(x, p,t,w;x',p',t',w')$, in which the Lagrangian manifold being projected is that representing the time evolution. This is so because in the extended, doubled space Berry's difficult time integral is subsumed under the same geometrical picture as all the other integrals. We hope to develop this picture in more detail in the future.

D. The Gutzwiller trace formula

We obtain the trace formula of Gutzwiller by reconsidering the trace of the unitary time-evolution operator, this time in the case that the Hamiltonian is conserved, but no other conserved quantities exist. It often happens in systems with these properties that all the periodic orbits are isolated, and we will assume that this is the case. We will set $t' = 0$ and write $\mathrm{Tr}\, U(t) = \mathrm{Tr}[I^\dagger U(t)]$ for the desired trace.

The manifolds L_U and L_I intersect in periodic orbits of period t, i.e., each point (x,p,x',p') on the intersection satisfies $(x,p) = (x',p')$, where (x,p) is on a periodic orbit of period t in the single phase space. Since the periodic orbits are isolated, the intersections are one dimensional. This situation is illustrated in Fig. 10, although, as mentioned in Sec. II, the picture is misleading in that it suggests that the intersection is stable under small perturbations (it is not). In-

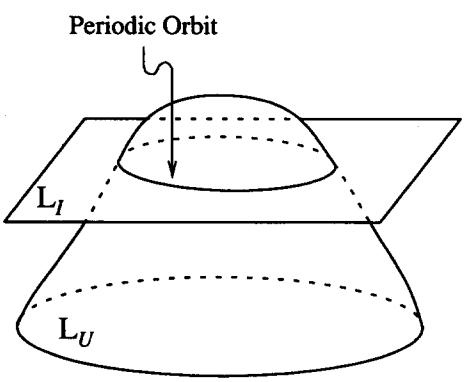

FIG. 10. In the Gutzwiller trace formula, the Lagrangian manifold in the doubled phase space representing the time evolution and that representing the identity canonical transformation intersect in the periodic orbits. These are the caustics of the system. The one-dimensional character of the intersection is not stable with respect to small perturbations in phase space.

deed, it seems clear that the very existence of this unstable configuration is due to a symmetry of the system, in this case, symmetry under time displacements. Other symmetries would have similar effects.

The complete set of commuting observables A and B of Eqs. (4.18) and (4.31) specify the manifolds L_I and L_U as before, but are not directly suitable for computing the amplitude determinant. Instead, we must follow the logic leading to Eq. (2.27), because L_I and L_U intersect in a one-dimensional curve. This requires us to find a new set of A's and B's.

To begin, we notice that the quantity $H(x,p) - H(x',p')$ generates orbits in the doubled phase space (with the doubled Poisson bracket being used to create Hamilton's equations) in which both (x,p) and (x',p') follow ordinary orbits in their respective phase spaces. Thus if an initial point (x,p,x',p') of the doubled phase space is such that (x,p), (x',p'), regarded as two points in the single phase space, lie on the same orbit with a given elapsed time between them, then they stay on this same orbit forever with the same time difference. This means that if an initial point (x,p,x',p') of the doubled phase space lies on L_I, it stays on L_I; and similarly if it lies on L_U. Therefore, $H(x,p) - H(x',p')$ generates displacements along the intersection of L_U with L_I, and may be identified with the functions $A_2 = B_2$ used in Eq. (2.27). Furthermore, the variable α_2 of Eq. (2.27) can be identified with the elapsed time along the periodic orbit, so the integral is just the time required to go around the intersection of L_I with L_U. This is not necessarily the period t of the orbit, because the orbit may not be primitive; instead, we have

$$\int_{L_U \cap L_I} d\alpha_2 = T = \frac{t}{n}, \tag{4.39}$$

where T is the primitive period and n is the number of iterations of the primitive period in time t.

Next, we must find $2f-1$ further A's and $2f-1$ further B's that commute with one another and which are constant on the manifolds L_I and L_U, respectively. The pattern established in Eqs. (4.18) and (4.31) suggests that we let the A's be the difference between the old and new variables in a canonical coordinate system on the single phase space in which H is one of the coordinates. We write (η,τ,E) for these canonical coordinates, where $\eta = (y,p_y)$ are $2f-2$ canonical variables in a surface of section, where τ is the elapsed time relative to the surface of section, measured along an orbit, and where E is interchangeable with H. We then take

$$\begin{aligned} A_\eta(\eta,\tau,E,\eta',\tau',E') &= \eta - \eta', \\ A_\tau(\eta,\tau,E,\eta',\tau',E') &= \tau - \tau', \\ A_E(\eta,\tau,E,\eta',\tau',E') &= E - E', \end{aligned} \tag{4.40}$$

where the first two equations specify the $2f-1$ components of A_1, and the last is A_2. These A's commute under the doubled Poisson bracket, and vanish on L_I.

Before proceeding to the B's, some comments on the coordinate system (η,τ,E) are in order. These coordinates were used by Berry[4] to compute the scars of Wigner functions in phase space. Berry did not point out, however, that these coordinates are not canonical unless the surface of section variables η in one energy surface are chosen in a particular way relative to the η's in nearby energy surfaces. In particular, the surface of section variables commonly employed in mechanical problems are generally not canonical on the full phase space. The problem is the Poisson bracket $\{\eta,\tau\}$, which does not vanish unless the η's are chosen to be constant along the τ trajectories. The fact that it is possible to choose η's that are simultaneously constant along both the τ trajectories and the H trajectories, i.e., so that $\{\eta,\tau\} = \{\eta,H\} = 0$, is due to the commutativity of the τ flow and the H flow. This in turn is a consequence of the relation $\{\tau,H\} = 1$. These facts are a straightforward consequence of Darboux's theorem,[18,40] in which H is initially chosen as one of the canonical coordinates.

In our analysis we will assume that the constructive program of Darboux's theorem has been followed, so that (η,τ,E) are canonical. We will also let the surface $\tau = 0$, which intersects the energy surfaces $H = E$ in the surfaces of section, serve as a branch cut for $\eta(x,p)$ and $\tau(x,p)$, so that these functions are single valued on phase space. Note that $\tau(x,p)$ is time independent, even though it signifies elapsed time.

It is also useful to consider branches of the functions η and τ other than the principal one. In particular, let $\eta = F(\eta',E')$ specify (say) the mth return map in the surface of section, and let $T_{\text{ret}}(\eta',E')$ be the corresponding return time. Then we introduce the alternative branches of η and τ, given by

$$\begin{aligned} \hat{\eta}(\eta,E) &= F(\eta,E), \\ \hat{\tau}(\eta,\tau,E) &= \tau - T_{\text{ret}}(\eta,E), \end{aligned} \tag{4.41}$$

which are the branches one obtains, not by following an orbit from a given point of phase space back to the most previous intersection with the surface of section, but rather back to the most previous and then forward in time m further intersections. The variables $(\hat{\eta},\hat{\tau},E)$ are also canonical variables on phase space, because of the preservation of Poisson brackets under Hamiltonian flows.

We may now write down the B's, in obvious analogy to Eq. (4.18). We have

$$\begin{aligned} B_\eta(\eta,\tau,E,\eta',\tau',E') &= \eta - \hat{\eta}(\eta',E'), \\ B_\tau(\eta,\tau,E,\eta',\tau',E') &= \tau - \hat{\tau}(\eta',\tau',E') - t, \\ B_E(\eta,\tau,E,\eta',\tau',E') &= E - E'. \end{aligned} \tag{4.42}$$

The first $2f-1$ of these constitute the B_1's and the last is B_2. These B's commute and have the property that every point on L_U satisfies $B = 0$ for some m, and conversely; if (x,p,x',p') is such a point, m is the number of times the orbit connecting (x',p') with (x,p) crosses the surface of section. Therefore, the intersections of L_I with L_U are given by $A = B = 0$ for some m, and points on one of these intersections are on a periodic orbit that crosses the surface of section m times. This m value must be divisible by n, the number of iterations of the primitive periodic orbit.

The amplitude determinant, which involves a $(2f-1)\times(2f-1)$ matrix of doubled Poisson brackets, may now be computed. Partitioning this matrix according to $2f-1 = (2f-2)+1$, we have

$$\{A_1,B_1\}_D=\begin{pmatrix}\Gamma\left(\mathbf{I}-\dfrac{\partial\tilde{F}}{\partial\eta}\right) & \Gamma\dfrac{\partial T_{\rm ret}}{\partial\eta}\\ -\dfrac{\partial\tilde{F}}{\partial E} & \dfrac{\partial T_{\rm ret}}{\partial E}\end{pmatrix}, \tag{4.43}$$

where the tilde represents the transpose and where Γ is the $(2f-2)\times(2f-2)$ constant matrix representing the cosymplectic form. This matrix is to be evaluated on the periodic orbit, where $\eta=\eta'$ and $E=E'$.

This matrix may be simplified. On the periodic orbit we have $\eta=F(\eta,E)$ and $t=nT=T_{\rm ret}(\eta,E)$, so we can eliminate η and solve for the primitive period T as a function of E. Taking differentials and solving for dT/dE, we have

$$n\frac{dT}{dE}=\frac{\partial T_{\rm ret}}{\partial E}+\frac{\partial\tilde{F}}{\partial E}\left(\mathbf{I}-\frac{\partial\tilde{F}}{\partial\eta}\right)^{-1}\frac{\partial T_{\rm ret}}{\partial\eta}. \tag{4.44}$$

This suggests that we multiply the first row of the matrix of Eq. (4.42) by the row vector $(\partial\tilde{F}/\partial E)(\mathbf{I}-\partial\tilde{F}/\partial\eta)^{-1}\Gamma^{-1}$ on the left, and add to the second row. We also write $\partial F/\partial\eta=\mathbf{M}^n$, where $\mathbf{M}$ is the linearized symplectic return map in the neighborhood of the primitive periodic orbit. The result is

$$\det\{A_1,B_1\}_D=n\frac{dT}{dE}\det(\mathbf{M}^n-\mathbf{I}). \tag{4.45}$$

It is now easy to write out the desired trace. The result is

$$\begin{aligned}\mathrm{Tr}\,U(t)&=\sum_n e^{-iE_nt/\hbar}\\&=\frac{1}{\sqrt{2\pi i\hbar}}\sum\frac{T\exp[(i/\hbar)R(t)-i\mu(\pi/2)]}{|n(dT/dE)|^{1/2}|\det(\mathbf{M}^n-\mathbf{I})|^{1/2}},\end{aligned} \tag{4.46}$$

where E_n are the energy eigenvalues. The second sum is taken over all periodic orbits of period t, for which $R(t)=nR(T)$ is Hamilton's principal function, evaluated around the orbit. This formula generalizes Tabor's result to the case that H is conserved. We note that in one degree of freedom, in which the surface of section is vacuous, the determinant factor in the denominator is simply replaced by unity.

Equation (4.46) is easily converted into a formula for the density of states, by performing the time integration of Eq. (3.12) by the stationary phase approximation, and then taking the imaginary part. In this way we obtain the Gutzwiller trace formula,

$$\rho(E)=\bar{\rho}(E)+\frac{1}{\pi\hbar}\sum\frac{T\cos[n(S(E)/\hbar)-\mu(\pi/2)]}{|\det(\mathbf{M}^n-\mathbf{I})|^{1/2}}, \tag{4.47}$$

where $\bar{\rho}(E)$ is the average density of states, where the sum is taken over all periodic orbits of energy E, and where $S(E)$ is the reduced action taken around the corresponding primitive periodic orbit. The fact that the Maslov index μ is proportional to the number of iterations of the primitive orbit is not obvious from this derivation (or from Gutzwiller's, either), but is proven by Robbins[22] and Creagh, Robbins, and Littlejohn.[23]

It would be more in accordance with the geometrical philosophy of this paper to examine the time integration in terms of Lagrangian manifolds in the extended, doubled phase space. We have not done this for several reasons, partly to avoid introducing another generalized phase space. Nevertheless, there can be no question that this is the space in which to fully understand the geometrical structure of the Gutzwiller trace formula. For example, the Maslov indices of the trace formula are ordinary Maslov indices on a Lagrangian manifold in this space, precisely in accordance with Maslov's general theory;[5] and the caustics that can occur in the trace formula are also best understood in this space.

A trivial yet interesting point about the transition from Eq. (4.46) to Eq. (4.47) by time integration is that the trace of the propagator (a complexified partition function) and the density of states are two representations of the same wave function, whose semiclassical expressions are supported by Lagrangian manifolds in the time-energy phase plane. The Lagrangian manifolds in question are the time-energy curves for the periodic orbits, and the periodic orbit sum can be interpreted as a sum over the branches of a WKB wave function. Interesting examples of these Lagrangian manifolds have been presented by Baranger and Davies.[41] These Lagrangian curves can also be regarded as slices through other Lagrangian manifolds in a phase space of higher dimensionality.

V. CONCLUSIONS

We will now conclude by commenting on the results presented and raising questions for further investigation.

First let us consider the results of Sec. II, in which the scalar product $\langle a|b\rangle$ is expressed in terms of intersections of Lagrangian manifolds in phase space. We have worked out the simplest case, in which the Lagrangian manifolds intersect transversally at isolated points. The next simplest case would be the one in which the Lagrangian manifolds intersect in partial or complete tangency, but still at isolated points. This case leads to the standard theory of caustics and catastrophes,[17] but it would be interesting to see the geometrical elements involved expressed in terms of the A and B foliations. As discussed in Sec. IV, this case would be useful for providing new insights into the caustic structure of Wigner functions, a matter of some interest recently, since Berry[4] has shown that the scars of periodic orbits in phase space are precisely such caustics.

Another case, which we did develop in Sec. II, is the one in which the A and B manifolds intersect in surfaces of dimensionality greater than zero. Here we assumed that there existed some coordinate transformations, replacing both the A's and B's by functions of themselves, such that some number of the new A's would coincide with the same number of new B's. This assumption leads to a simplification of the computation of the scalar product $\langle a|b\rangle$, in that the integrand becomes independent of α_2, as discussed in the derivation of Eq. (2.27). Although this assumption makes the case we considered a rather special one, it is the simplest case of higher dimensional intersections of Lagrangian manifolds, and it is also the one we need for the Gutzwiller trace formula. It also explains, in a sense, why Gutzwiller is able to do the time integral around the periodic orbit, i.e., why the integrand is independent of time.

But it is not the most general case of higher dimensional intersections of Lagrangian manifolds, as counterexamples will show. Therefore, the question is raised, what is the deeper meaning of the existence of a coordinate transformation causing some A's and B's to coincide? The answer seems to be connected with symmetries, in the sense that any observable which can be expressed as a function of only the A's, or alternatively of only the B's, must commute with both of them; and the set of all such functions forms an Abelian group. No doubt the proper way to express the scalar product $\langle a|b\rangle$ of Eq.(2.27) would be in terms of some original collection of A's and B's, combined with the generators of the symmetry group. This would yield a much more elegant calculation of Tr $U(t)$, in the case of a time-independent system, than the one we presented in Sec. IV, in that it would not be necessary to work with surface of section coordinates. Instead, we could work with the original x's and p's, and the generator of the symmetry group, $H(x,p) - H(x',p')$.

Perhaps the most interesting result of Sec. III, as simple as it is, is the realization of the close connection between the Green's function and the propagator in the surface of section. This shows, in a sense, why the surface of section monodromy matrix occurs in Gutzwiller's formula, while that for the full phase space occurs in Tabor's. In fact, apart from the period T of the orbit in Gutzwiller's formula and the reduction of the dynamics by one degree of freedom, the two formulas are identical. The factor T can be explained as arising from the fact that the surface of section evolution is really governed by H, the real Hamiltonian, and not by the Hamiltonian F for the surface of section evolution. This fact also explains why Gutzwiller's trace formula cannot be represented as a sum over the quasiphases of the nominal quantized surface of section mapping (a tempting idea which was not mentioned in Sec. III because it fails).

A natural extension of the work of Sec. IV would be to include the effects of arbitrary symmetry groups. The extreme cases of complete integrability and complete chaos are well known, but intermediate cases are less so. Non-Abelian symmetries, such as rotation, would be interesting to incorporate into the geometrical framework presented here; this would build on the work of Strutinskii and Magner,[42] but operating from a rather different standpoint. Discrete symmetries in the trace formula have already been dealt with by Robbins.[43] There are also certain improvements that can be made in the elimination of degrees of freedom from the amplitude determinant of the trace formula, when symmetries exist. Finally, one should work in an extended phase space, in order to geometrize the time integrations and place them on an equal footing with all the others. These and other issues will be considered in the future.

ACKNOWLEDGMENTS

I am especially grateful to the Institute for Theoretical Physics at the University of California, Santa Barbara, and to its director, Bob Schrieffer, for hosting me while much of this work was done; and to Eric Heller and Martin Gutzwiller for organizing the Workshop on Small Quantum Systems at the Institute. I am also grateful for useful discussions with many people, including Nandor Balazs, Stephen Creagh, Shmuel Fishman, Eric Heller, Martin Gutzwiller, Marcos Saraceno, André Voros, and Michael Wilkinson, and to Alan Weinstein for providing me with references on the theory of Fourier integral operators. I would also like to thank Bill Fine and the Theoretical Physics Institute of the University of Minnesota for further hospitality and support during the completion of this work.

This work was supported by the National Science Foundation under Grant Number NSF-PHY82-17853, supplemented by funds from NASA, by the Office of Energy Research, Office of Basic Energy Sciences, of the U.S. Department of Energy under Contract Number DE-AC03-76SF00098, by the National Science Foundation under Grant Number NSF-PYI-84-51276, and by the Theoretical Physics Institute of the University of Minnesota.

[1] M. C. Gutzwiller, J. Math. Phys. **8**, 1979 (1967); **10**, 1004 (1969); **11**, 1791 (1970); **12**, 343 (1971).
[2] R. Balian and C. Bloch, Ann. Phys. **60**, 401 (1970); **63**, 592 (1971); **64**, 271 (1971); **69**, 76 (1972); **85**, 514 (1974).
[3] M. Tabor, Physica D**6**, 195 (1983).
[4] M. V. Berry, Proc. R. Soc. London Ser. A **423**, 219 (1989).
[5] V. P. Maslov and M. V. Fedoriuk, *Semi-Classical Approximation in Quantum Mechanics* (Reidel, Boston, MA, 1981).
[6] M. V. Berry and M. Tabor, Proc. R. Soc. London Ser. A **349**, 101 (1976); J. Phys. A **10**, 371 (1977).
[7] M. V. Berry, in *Chaotic Behaviour in Deterministic Systems*, edited by G. Iooss, R. G. Helleman, and R. Stora (North Holland, Amsterdam, 1983), p. 171; in *Proceedings of the Les Houches School on Chaos and Quantum Physics*, session No. 52, August 1989 edited by M. J. Giannoni and A. Voros (to be published by North Holland).
[8] A. M. Ozorio de Almeida, *Hamiltonian Systems: Chaos and Quantization* (Cambridge U. P., Cambridge, 1988).
[9] N. L. Balazs and A. Voros, Phys. Rep. **143**, 109 (1986).
[10] W. H. Miller, Adv. Chem. Phys. **25**, 69 (1974).
[11] R. Abraham and J. Marsden, *Foundations of Mechanics* (Benjamin/Cummings, Reading, MA, 1978), 2nd ed.
[12] L. Hörmander, Acta Math. **127**, 79 (1971).
[13] J. J. Duistermaat and V. W. Guillemin, Invent. Math. **29**, 39 (1975).
[14] R. J. Blattner and B. Kostant, in *Géométrie Symplectique et Physique Mathématique*, Proceedings of the Colloques Internationaux du Centre National de la Recherche Scientifique, Aix-en-Provence, No. 237 (Éditions du CNRS, Paris, 1975).
[15] A. Weinstein, *Lectures on Symplectic Manifolds*, Conference Board of the Mathematical Sciences No. 29 (American Mathematical Society, Providence, RI, 1976).
[16] V. Guillemin and S. Sternberg, *Geometric Asymptotics* (American Mathematical Society, Providence, RI, 1977).
[17] M. V. Berry and C. Upstill, in *Progress in Optics*, edited by E. Wolf (North-Holland, Amsterdam, 1980), Vd. XVIII, p. 250.
[18] V. I. Arnold, *Mathematical Methods of Classical Mechanics* (Springer, New York, 1978).
[19] I. C. Percival, Adv. Chem. Phys. **36**, 1 (1977).
[20] J. B. Delos, Adv. Chem. Phys. **65**, 161 (1986).
[21] L. D. Landau and E. M. Lifshitz, *Mechanics* (Pergamon, New York, 1960).
[22] J. M. Robbins, "Maslov Indices in the Gutzwiller Trace Formula," to be published in Nonlinearity.
[23] S. Creagh, J. M. Robbins, and R. G. Littlejohn, Phys. Rev. A **42**, 1907 (1990).
[24] M. V. Berry and N. L. Balazs, J. Phys. A **12**, 625 (1979); M. V. Berry, N. L. Balazs, M. Tabor, and A. Voros, Ann. Phys. **122**, 26 (1979).
[25] J. H. Van Vleck, Proc. Natl. Acad. Sci. USA **14**, 178 (1928).
[26] L. S. Schulman, *Techniques and Applications of Path Integration* (Wiley, New York, 1981).
[27] M. V. Berry and K. E. Mount, Rep. Prog. Phys. **35**, 315 (1972).
[28] J. B. Keller, in *Calculus of Variations and its Applications*, edited by L. M. Graves (McGraw-Hill, New York, 1958).
[29] E. J. Heller, Phys. Rev. Lett. **53**, 1515 (1984).
[30] M. Saraceno, Ann. Phys. (NY) **199**, 37 (1990).

[31] A. Voros, J. Phys. A **21**, 685 (1988).
[32] M. C. Gutzwiller, Physica D**5**, 183 (1982).
[33] V. I. Man'ko, in *Lie Methods in Optics*, edited by J. S. Mondragón and K. B. Wolf (Springer, New York, 1986), p. 193.
[34] W. H. Miller, J. Phys. Chem. **83**, 960 (1979).
[35] J. M. Robbins, S. Creagh, and R. G. Littlejohn, Phys. Rev. A **39**, 2838 (1989).
[36] R. G. Littlejohn, Phys. Rep. **13**, 193 (1986).
[37] H. Goldstein, *Classical Mechanics* (Addison-Wesley, Reading, MA, 1980), 2nd ed.
[38] A. M. Ozorio de Almeida and J. H. Hannay, J. Phys. A **20**, 5873 (1987).
[39] M. V. Berry, Philos. Trans. R. Soc. London **287**, 237 (1977).
[40] R. G. Littlejohn, J. Math. Phys. **20**, 2445 (1979).
[41] M. Baranger and K. T. R. Davies, Ann. Phys. **177**, 330 (1987).
[42] V. M. Strutinskii and A. G. Magner, Sov. J. Part. Nucl. **7**, 138 (1976).
[43] J. M. Robbins, Phys. Rev. A **40**, 2128 (1989).

How solvable is (2+1)-dimensional Einstein gravity?

Vincent Moncrief
Department of Physics and Department of Mathematics, Yale University, 217 Prospect Street, New Haven, Connecticut 06511

(Received 7 May 1990; accepted for publication 25 July 1990)

In this paper, the relationship between Witten's approach to the (2 + 1)-dimensional, vacuum Einstein equations (for spatially compact space-times) and the conventional Annowitt, Deser, and Misner (ADM) Hamiltonian approach is discussed. It is argued (especially for the space-times with higher genus Cauchy surfaces) that *neither approach is complete in itself*; Witten's because it does not provide a technique (even at the classical level) for recovering the space-time metric and the conventional approach because it provides no mechanism for solving a seemingly intractable set of Hamilton equations. It is also argued, however, that the two formulations are instead *complementary* in the sense that the Wilson loops, which play a key role in Witten's approach, provide (at least in principle) a mechanism for solving the reduced Hamilton equations and thereby completing the picture at the classical level. An example of this synthesis for the (explicitly computable) case of genus-one hypersurfaces is provided. The more tenuous problem of whether this synthesis can be extended to the quantized Einstein equations will also be discussed. A principal open question is whether the Wilson loops, when expressed in terms of the ADM canonical variables, can be ordered in such a way as to preserve, quantum mechanically, their (classical) Poisson bracket algebra.

I. INTRODUCTION

In a recent paper, Witten has argued that the (2 + 1)-dimensional, vacuum Einstein equations are an exactly solvable system both classically and quantum mechanically.[1] At first glance the result sounds obvious since, in 2 + 1 dimensions, Einstein's equations imply that space-time is flat. Thus there seem to be no gravitational degrees of freedom left to study. On manifolds of the form $\Sigma \times R$ however, where Σ is a compact, orientable surface, there are *topological degrees of freedom* to consider and Witten's conclusion is far from evident. In fact, if one attempts to study the dynamics of these topological variables (which can be thought of as coordinates on the cotangent bundle of the Teichmüller space associated to Σ) by conventional Hamiltonian methods one finds (at least in the more interesting cases for which the genus g of Σ is greater than or equal to 2), a seemingly intractable Hamiltonian governing the dynamics.[2]

By contrast, Witten finds that the classical solutions can be labeled by a set of functions (which can also be thought of as coordinates for the cotangent bundle of Teichmüller space) that have *no dynamics at all*, i.e., which are constants of the motion for the dynamics determined by Einstein's equations. Of these 12 g–12 independent coordinate functions (when $g \geqslant 2$), Witten chooses a Poisson commuting subset of 6 g–6 functions that determine a real polarization on the phase space and formulates the associated quantum Hilbert space as the space of square integrable, complex functions of these 6 g–6 commuting coordinates. Since both the coordinates and their conjugate momenta are constants of the motion in this formulation, the Hamiltonian is identically zero and thus there is no Schrödinger equation or Heisenberg equations of motion to consider. The dynamics are *trivial* both classically and quantum mechanically.

There is a rigorous and coordinate independent sense (at the classical level) however in which the dynamics of these solutions are genuinely *nontrivial*. One can prove (by the same methods discussed in Ref. 3 for the vacuum Einstein equations in 3 + 1 dimensions) that none of the solutions discussed in Ref. 2 (with $g \geqslant 2$) are stationary (i.e., admit a globally defined timelike Killing field) and that only a 6 g–6-dimensional subfamily of the 12 g–12-dimensional space of these classical solutions are homothetically stationary (i.e., admit a globally defined timelike, homothetic Killing field). This latter set is well known and can be recovered, as also noted by Witten, by taking the quotient of three-dimensional Minkowski space with respect to suitably chosen discrete subgroups of the Lorentz group. The dynamics in this case are a relatively simple "rescaling" that reflects the homothetic symmetry of these solutions. These particular solutions can easily be recovered from the Hamiltonian analysis of Ref. 2 as well. The highly nontrivial evolution of the canonical coordinates for the remaining solutions, determined in principle by the solution of the Hamilton equations, reflects the nonstationarity of these solutions (in even the homothetic sense).

How can one reconcile the simple picture provided by Witten with the dramatically more complicated conventional point of view, or is there a genuine conflict in the two analyses? An answer begins to emerge when one looks more carefully at the constants of the motion that Witten uses to coordinatize the classical phase space. These constants of the motion arise as "Wilson loops" for a connection of an ISO(2,1) bundle over space-time that Witten introduces in order to reformulate the Einstein equations as an effective "Chern–Simons gauge theory." One can evaluate the Wilson loops, for example, on a spacelike hypersurface diffeomorphic to Σ and then discover that the results of the computation are invariant under deformations of Σ as a consequence of the flatness of the ISO(2,1) connection (which corresponds, in the Chern–Simons formulation, to the flatness of the space-time). Thus the Wilson loops, which provide the

 0022-2488/90/122978-05$03.00

topological information needed to characterize the solutions, are invariant under time translation no matter what gauge one chooses to represent "time." This is so in spite of the fact that there is a genuine evolution of the geometrical properties of the hypersurfaces taking place, under such time translation, which reflects the nonstationarity of the space-times being considered. In general, if one deforms Σ to some Σ' the first and second fundamental forms induced on these hypersurfaces by the space-time metric changes accordingly and are determined in principle (for the special case in which Σ and Σ' are hypersurfaces of constant mean curvature) by the solution of the Hamilton equations discussed in Ref. 2.

But the Wilson loops, by their very constancy, are insensitive to this deformation and therefore cannot, by themselves, distinguish any one hypersurface in the chosen space-time from any other. They may label the chosen space-time unambiguously (and in a way which is hypersurface independent) but they give, by themselves, no information about the (equally invariantly defined) evolution which is rigorously known to be taking place. To recover the space-time metric however, a knowledge of this evolution is unavoidable since that is exactly what the metric itself must describe. Indeed Witten's paper offers no examples (other than the special ones with homothetic symmetry and therefore trivial evolution in even the Hamiltonian sense) of metrics that satisfy Einstein's equations on $\Sigma \times \mathbf{R}$ and provides no direct method for their construction except to say that they can in principle be obtained by taking quotients of Minkowski space by suitably chosen discrete subgroups of the Poincaré group.

What seems unsatisfying about this state of affairs is that there are interesting dynamical questions about the class of space-times under study which neither approach seems capable of answering. In particular one could ask whether the classical solutions have any tendency to "change topology" in the sense that, under evolution, their genus g hypersurfaces evolve to "pinch off" toward something like a connected sum of lower genus surfaces or whether, conversely, the solutions could exhibit a "global existence" property which excludes such singular behavior over the maximal allowed range of the mean curvature time variable τ [i.e., over the range of either $\tau \in (0, \infty)$ or $(-\infty, 0)$]. These are interesting open questions which neither approach seems at present able to answer; Witten's because his variables are insensitive to the evolution altogether and the conventional Hamiltonian approach because of the apparent intractability of the Hamiltonian.

The main point of this paper is to suggest that there may be a well-defined way of answering such questions and, at the same time, of reconciling the disparate viewpoints described above. The basic idea is simply *to evaluate the Wilson loops in terms of the conventional Hamiltonian variables.* This can always be done (at least in principle) since, on the one hand, the Hamiltonian variables completely determine the metrical properties of the space-time relative to the constant mean curvature time slicing (and modulo spatial coordinate gauge transformations) and since, on the other, the Wilson loops are invariant with respect to local Lorentz and coordinate gauge transformations (and thus pass to the quotient space parametrized by the Hamiltonian variables). The resulting mapping connecting the $12g-12$ Hamiltonian variables and the $12g-12$ independent Wilson loops cannot simply be a fixed (local) diffeomorphism of the phase space but must incorporate the mean curvature variable τ which plays the role of time. Otherwise, one could simply solve for the Hamiltonian variables and discover that they are necessarily constants of the motion—a conclusion which is known to be *false.*

The idea however is to solve for the Hamiltonian variables in terms of the complete set of Wilson loops *and the time variable* τ and thus to derive the general solution of the otherwise seemingly intractable Hamilton equations without ever writing down the Hamiltonian itself. Thus given the general solution of Hamilton's equations one could retrace the steps of Ref. 2 in reverse order and (at least in principle) reconstruct the space-time metric corresponding to any particular such solution.

We have not yet shown that this procedure is always mathematically well defined in the higher genus cases. In particular, the transformation might fail to be smoothly invertible (even locally) but this seems rather unlikely. In the following section we shall derive the transformation explicitly for the one case where everything is explicitly computable—the genus-one case for which Σ is diffeomorphic to the two-torus. In this case we shall find that transformation is indeed smoothly invertible almost everywhere and reproduces the known exact solution for the $\mathbf{T}^2 \times \mathbf{R}$ space-times. As is well known, the solution curves for this problem coincide with geodesics on hyperbolic two-space (which is the associated "Teichmüller" space for $\mathbf{T}^2$) although τ does *not* coincide with affine parametrization of these geodesics.

Even in the absence of mathematical subtleties however the transformation suggested above can probably not be given explicitly in the higher genus cases. The reason for this is that the evaluation of the Wilson loops in terms of the Hamiltonian Cauchy data apparently requires some highly nontrivial mathematical operations such as the solution of the Lichnerowicz equation for the conformal factor that implements the Hamiltonian constraint. On the other hand it seems plausible that a *proof* of the smooth invertibility (almost everywhere) of this transformation might be possible and that such a proof would provide an affirmative answer to the "global existence" conjecture mentioned above and thereby exclude the possibility of "topology change" for the classical solutions throughout the maximal allowed range of mean curvature (i.e., the range which exhausts the maximal Cauchy developments of the space-times under study).

The above remarks have mostly concerned the solution of the *classical* Einstein equations but Witten's paper also deals with their quantum counterpart and a proposal for solving $(2+1)$-dimensional *quantum gravity* by means of Wilson loops as well. At the classical level our point of view has been that the Wilson loops by themselves are incomplete (except as *labels* for the classical solutions) and that a full solution of the classical Einstein equations (which, at least in principle, entails the reconstruction of the space-time metrics) requires the computation of these conserved quantities

in terms of more conventional Hamiltonian Cauchy data. For similar reasons, one might question whether Witten's quantum solution is satisfying as it stands.

His proposal is to choose a Poisson commuting subset of 6 g–6 independent Wilson loops (which can be thought of as coordinates on a space diffeomorphic to Teichmüller space) and to define the quantum Hilbert space as the space of complex-valued square integrable functions of these "canonical coordinates." Since both these Wilson loops and their conjugate partners are constants of the (classical) motion the Hamiltonian is identically zero and thus there is no quantum Schrödinger equation or Heisenberg equations of motion to consider. This is the quantum analog of the fact that, as far as the Wilson loops themselves are concerned, there are no Hamilton equations to consider in the classical theory.

From our point of view, however, the constancy of the Wilson loops is analogous to the fact that any classical Hamiltonian system may be transformed by a *time-dependent* canonical transformation to a new set of canonical variables (via Hamilton–Jacobi theory, for example) which are all constants of the motion. From this point of view Witten's quantization procedure would seem to correspond to first expressing the classical problem in terms of variables for which the Hamiltonian is identically zero (the usual objective of Hamilton–Jacobi theory) and then "quantizing" the transformed classical problem by simply declaring that it's quantum Hilbert space is the space of complex square integrable functions of a complete commuting subset of the new canonical variables.

By contrast to this procedure one could adopt the more conventional approach to quantization by taking the reduced Hamiltonian (which is nonzero) and formulating, after a suitable choice of operator ordering, either the Schrödinger equation or the Heisenberg equations of motion. For the example to be discussed below this quantization is perfectly tractable since, for example, the eigenstates of the associated Schrödinger operator are simply the eigenfunctions of the Laplacian on hyperbolic two-space. For the higher genus cases, however, this approach seems completely intractable since, as we have mentioned, even the classical expression for the Hamiltonian is not known explicitly.

Perhaps, however, there is a quantum analog to the technique we described above for solving the classical Hamilton equations by means of the Wilson loops. The idea would be first to try to choose an operator ordering in the quantized expressions for the Wilson loops that preserves (as far as possible) the Poisson bracket algebra of these constants of the motion. One could then try to regard the resulting expressions as an implicit solution of the associated Heisenberg equations of motion which, in principle, determines the conventional canonical coordinate and momentum operators as functions of certain constant operators and the time variable τ.

Given the implicit nature of the Wilson loop expressions (in the higher genus cases) there is no guarantee that this program can be made to work. It is interesting to note however that, in the context of spatially open space-times with point particle sources, Martin has been able to prescribe an ordering of the quantized Wilson loops that preserves their classical commutation relations.[4] He is working however with Witten's original unreduced canonical variables (the orthonormal frame and associated Lorentz connection) in terms of which the Wilson loops are always linear in the canonical momenta (i.e., in the frame fields). As we shall see in the example given below the reduced expressions for the Wilson loop integrals no longer have this simple linearity property (when expressed in terms of the conventional canonical variables) so it is not clear whether Martin's result can be preserved upon reduction for the quantum problems considered here.

II. FLAT SPACE-TIMES ON $T^2\times R$ VIA WILSON LOOPS

As discussed in Ref. 2 the flat metrics on $T^2\times R$ are spatially homogeneous and can be expressed in suitable coordinates (for which x^1 and x^2 are "angle" coordinates on S^1, each defined mod 2π) in the form:

$$ds^2 = -N(t)^2\,dt^2 + e^{2\mu(t)}(dx^1)^2 + e^{2\nu(t)}(dx^2+\beta(t)dx^1)^2. \tag{1}$$

To compute the Wilson loops we introduce the orthonormal frame:

$$e^{(0)} = N(t)dt, \quad e^{(1)} = e^{\mu(t)}\,dx^1,$$
$$e^{(2)} = e^{\nu(t)}(dx^2+\beta(t)dx^1) \tag{2}$$

and compute the connection one-forms $\omega_{(a)(b)} = \omega_{(a)(b)\mu}\,dx^\mu$,

$$\begin{aligned}\omega_{(1)(0)} &= -\omega_{(0)(1)} = A(t)dx^1 + C(t)dx^2,\\ \omega_{(2)(0)} &= -\omega_{(0)(2)} = B(t)dx^1 + D(t)dx^2,\\ \omega_{(1)(2)} &= -\omega_{(2)(1)} = \tfrac{1}{2}e^{\nu(t)-\mu(t)}\beta_{,t}\,dt,\end{aligned} \tag{3}$$

where

$$\begin{aligned}A(t) &= [1/N(t)](e^{\mu}\mu_{,t} + \tfrac{1}{2}e^{2\nu-\mu}\beta\beta_{,t}),\\ B(t) &= [1/N(t)](\beta e^{\nu}\nu_{,t} + \tfrac{1}{2}e^{\nu}\beta_{,t}),\\ C(t) &= [1/N(t)](\tfrac{1}{2}e^{2\nu-\mu}\beta_{,t}),\\ D(t) &= [1/N(t)](e^{\nu}\nu_{,t}).\end{aligned} \tag{4}$$

Since the components of these one-forms are independent of position on the $t=$ constant hypersurfaces it is relatively easy to compute the Wilson loops for simple closed curves of the form $x^2=$ constant (the "a-loops") or $x^1=$ constant (the "b loops") or $x^1=(p/q)\,x^2$ ("twisting loops") by applying the techniques used by Martin in Ref. 4. The twisting loops do not give additional independent conserved quantities but instead give certain functions of the conserved quantities determined by the a and b loops. Nevertheless, it is of interest to carry them along as though they were independent quantities. From the first representation of ISO(2,1) discussed by Martin one gets the conserved quantities

$$C_1 = A(t)^2 + B(t)^2, \quad C_2 = C(t)^2 + D(t)^2, \tag{5}$$

from the a and b loops, respectively, and

$$C_3 = A(t)C(t) + B(t)D(t), \tag{6}$$

from the twisting loops. From the second representation discussed by Martin one gets

$$C_4 = e^{\mu(t)}B(t) - e^{\nu(t)}\beta(t)A(t),$$
$$C_5 = e^{\nu(t)}C(t), \tag{7}$$

from the a and b loops, respectively, and

$$C_6 = e^{\mu(t)}D(t) - e^{\nu(t)}A(t) - e^{\nu(t)}\beta(t)C(t) \tag{8}$$

from the twisting loops. It is straightforward to verify directly that C_1–C_6 are conserved by Einstein's equations but this will follow more immediately after passing to the Hamiltonian formulation.

It is convenient to define the new variables

$$q^1 = \nu - \mu, \quad q^2 = \beta, \quad q^3 = \nu + \mu \tag{9}$$

and to introduce their conjugate momenta p_1, p_2, p_3 in the usual way,

$$\sum p_i q^i_{,t} = \sum \pi^{ab} g_{ab,t}, \tag{10}$$

where $\{g_{ab},\pi^{ab}\}$ are the conventional Arnowitt, Deser, and Misner (ADM) canonical variables. In these variables q^1 and q^2 parametrize the conformal metric

$$h_{ab} = g_{ab}/\sqrt{^{(2)}g}, \tag{11}$$

while q^3 parametrizes the volume element and the mean curvature τ is given by

$$\tau = p_3/e^{q^3}. \tag{12}$$

As discussed in Ref. 2 (but with a different normalization of the time variable chosen to avoid the factors of 2π occurring there) we can solve the Hamiltonian constraint for the spatial volume element and impose the temporal coordinate condition $\tau = e^t$ to obtain the reduced ADM Hamiltonian for the unconstrained variables $\{q^1,q^2,p_1,p_2\}$. The result is

$$H = \sqrt{(p_1)^2 + e^{-2q^1}(p_2)^2}, \tag{13}$$

which can be viewed as the square root of the usual generator for geodesic motion on hyperbolic two-space which, in these coordinates, is simply R^2 equipped with the metric

$$dl^2 = (dq^1)^2 + e^{2q^1}(dq^2)^2. \tag{14}$$

Since H is conserved it generates the same solution curves as H^2, but relative to a nonconventional affine parameter (whose relation to the usual one depends on the value of H).

We can now express the conserved quantities determined by the Wilson loops in terms of the unconstrained canonical variables. The result is

$$\begin{aligned} C_1 &= \tfrac{1}{2}\tau e^{-q^1}\{(H-p_1)(1+(q^2)^2e^{2q^1}) \\ &\quad -2(q^2p_2 - p_1)\}, \\ C_2 &= \tfrac{1}{2}\tau e^{q^1}(H-p_1), \\ C_3 &= \tfrac{1}{2}\tau e^{q^1}\{q^2(H-p_1) - p_2 e^{-2q^1}\}, \\ C_4 &= \tfrac{1}{2}\{p_2 e^{-2q^1} + 2q^2p_1 - p_2(q^2)^2\}, \\ C_5 &= \tfrac{1}{2}p_2, \quad C_6 = p_1 - q^2p_2. \end{aligned} \tag{15}$$

As mentioned previously these are not all independent but in fact satisfy

$$H^2 = (C_6)^2 + 4C_4C_5, \quad (C_3)^2 = C_1C_2. \tag{16}$$

It is not difficult to check that C_4, C_5, and C_6 are the usual conserved quantities associated with the SL(2, R) isometry group of hyperbolic two-space; one can read off the Killing fields as coefficients of the canonical momenta. The first three quantities, C_1, C_2, and C_3, depend explicitly upon the time through the factor $\tau = e^t$, but it is straightforward to check that their Poisson brackets with H cancel their derivatives with respect to the explicit dependence on t and thus that they are conserved as expected.

Although it is far from obvious from their expressions in terms of the ADM variables, it is also straightforward to check that C_1, C_2, and C_3 *Poisson commute* with one another. In Witten's approach to quantization one would take two of these variables, say C_1 and C_2, as independent canonical coordinates and express the physical states as time independent, square integrable, complex-valued functions of these quantities. In the more conventional approach to quantization one would construct the operator analog of H (which, in this case, would be the square root of the negative of the usual covariant Laplacian on hyperbolic two-space) and formulate either the Schrödinger equation or Heisenberg equations of motion in terms of it. It seems far from clear whether there is any precise (e.g., unitary) connection between these two different "quantizations."

From the classical point of view however there is no inconsistency in the different formulations. Indeed one can simply choose four independent conserved quantities, C_1, C_2, C_4, and C_5, for example, set them equal to constant values and solve the resulting set of simultaneous equations for the ADM canonical variables as functions of the chosen constants and the time variable τ. In this way one solves the Hamilton equations by purely algebraic manipulations. The algebra can be carried out explicitly for the present problem but is slightly tedious since there are several sign choices to be considered in getting all the different roots and since the chosen conserved quantities are not globally independent on phase space (although they are independent almost everywhere). It is of interest to note however that if one computes the Jacobian determinant of the transformation expressing $\{C_1,C_2,C_4,C_5\}$ in terms of $\{q^1,q^2,p_1,p_2\}$ the result is simply $-(C_3)^2$ and is thus constant along solution curves. Thus the chosen Wilson loop constants fail to provide a good coordinate system for the space of solutions only for those particular solutions characterized by $C_3 = 0$. It seems likely that one could "patch the seams" in this coordinate system by making a different choice of Wilson loop constants to cover the solutions excluded in the choice above but we shall not pursue that issue here.

Even though the conventional approach to quantization is perfectly tractable for the present problem it is of interest to ask whether one could sidestep the direct solution of Schrödinger's equation but arrive at an equivalent quantum mechanical solution by using a suitably quantized form of the Wilson loop constants of motion to solve the Heisenberg equations of motion. This would be the direct quantum analog of the method used above to solve Hamilton's equations for the classical problem. To see what one might mean by a "suitably quantized form" of the Wilson loop constants let us first consider the complete algebra of their Poisson brack-

ets in the classical theory. It is straightforward to verify that if one defines

$$P_0 = \tfrac{1}{2}(C_1 + C_2), \quad P_1 = \tfrac{1}{2}(C_1 - C_2), \quad P_2 = C_3,$$
$$J_{12} = C_5 - C_4, \quad J_{01} = -C_6, \quad J_{02} = C_4 + C_5, \quad (17)$$

and identifies the $\{P_\mu\}$ as the translation generators and the $\{J_{\mu\nu}\}$ as the rotation and boost generators in three-dimensional Minkowski space then indeed the Poisson brackets of these quantities generate precisely the Lie algebra of ISO(2,1), the Poincaré group in three dimensions. Thus it seems clear that if one wishes to find an "inverse solution" of the Heisenberg equations by means of the Wilson loop constants he should seek an ordering of the canonical variables $\{q^1,q^2,p_1,p_2\}$ in the expressions for the generators $\{P_\mu,J_{\mu\nu}\}$ so as to preserve the Lie algebra of the Poincaré group.

We conclude this section by pointing out that the lack of independence of the constants of the motion $\{C_1-C_6,H\}$ pointed out above has an elegant expression in terms of the Poincaré group generators:

$$H^2 = (J_{01})^2 + (J_{02})^2 - (J_{12})^2, \quad -P_0{}^2 + P_1{}^2 + P_2{}^2 = 0. \quad (18)$$

III. CONCLUDING REMARKS

One often hears of the desirability of constructing a complete set of "observables" for the gravitational field, i.e., a maximal independent set of functions of the ADM canonical variables $\{g_{ab},\pi^{ab}\}$ which Poisson commute with all of the constraints and thus define a complete set of constants of the motion for Einstein's equations. Indeed this is a key step in the long-standing program of Bergmann and Komar to quantize gravity in 3 + 1 dimensions.[5] One aspect of Witten's work (as well as that of Ashtekar *et al.* in a closely related work based on the use of Ashtekar's variables[6]) has been to show that the Wilson loops provide a set of observables for the gravitational field in 2 + 1 dimensions.

A main point of this paper has been however that such observables, *by themselves*, merely provide an unambiguous set of *labels* for the space-times in question and do not yield, without further information, the geometrical properties of these space-times. In this respect they are somewhat analogous to the complete sets of initial positions and momenta that label the solutions of any problem in Hamiltonian mechanics. To understand the properties of the solutions themselves one needs to know the transformation which relates these labels to the conventional canonical variables.

The Wilson loops, however, are in principle expressible in terms of the conventional Cauchy data induced on an arbitrary slicing of space-time. If one expresses these loops in terms of the Cauchy data for the chosen slicing, he should obtain an implicit form of the solution of the reduced Einstein equations. From this point of view Witten's approach and the conventional Hamiltonian one are *complementary* to one another. Neither provides a full solution to the (classical) vacuum Einstein problem (in say the higher genus cases) but together they have the potential of solving this problem in at least an implicit way.

The connection between Witten's approach and the conventional one seems much more tenuous at the quantum level however. Here, to use the Wilson loops to solve the Heisenberg equations of motion implicitly, one would presumably need to order the operator expressions for these (classically) conserved quantities so that the quantum commutator algebra preserves the classical Poisson bracket algebra. Whether this can be done in general or whether the solution, if it exists, is unique are open questions. The results of Martin[4] and Carlip[7] on the somewhat related problem of particle scattering in spatially open geometries (via Wilson loops) may provide some insights in this direction as could a further development of the suggestions made in this paper or those of Ashtekar *et al.* in Ref. 6.

An interesting question about (2 + 1)-dimensional quantum gravity which we have not really touched upon is *whether general covariance is compatible with quantization.* The Hamiltonian approach described in Ref. 2 and used in the present paper works with a *rigidly fixed temporal gauge condition*, the constant mean curvature slicing, often called the York time gauge. In principle, one could have formulated the reduced Einstein equations in terms of an arbitrary slicing and appealed to the Wilson loop constants to solve the associated classical equations of motion. Quantum mechanically however it is far from evident that a corresponding solution exists (for the higher genus cases) in even a single such gauge, much less in an arbitrary one.

What, after all, is a *quantum space-time* and does such an object admit a representation in terms of different *space-like slicings*? These questions seem interesting and nontrivial even in the context of the $\mathbf{T}^2\times\mathbf{R}$ models studied here, not to mention the higher genus cases where an *explicit* formulation still seems out of reach.

ACKNOWLEDGMENTS

I have benefitted from numerous conversations with my Mathematics and Physics colleagues at Yale as well as from discussions with Abhay Ashtekar, Steven Carlip, Gary Horowitz, Viqar Husain, Karel Kuchar, Michael Ryan, and Lee Smolin.

This research was supported in part by NSF Grant PHY 8903939 to Yale University.

[1] E. Witten, Nucl. Phys. B **311**, 46 (1988).
[2] V. Moncrief, J. Math. Phys. **30**, 2907 (1989).
[3] D. Eardley, J. Isenberg, J. Marsden, and V. Moncrief, Commun. Math. Phys. **106**, 137 (1986).
[4] S. Martin, Nucl. Phys. B **327**, 178 (1989).
[5] See, for example the article "The Phase Space Formulation of General Relativity and Approaches Towards its Canonical Quantization," by P. Bergmann and A. Komar in *General Relativity and Gravitation, Vol. 1*, edited by A. Held (Plenum, New York, 1980).
[6] A. Ashtekar, V. Husain, C. Rovelli, J. Samuel, and L. Smolin, Class. Quantum Grav. **6**, L185 (1989).
[7] S. Carlip, Nucl. Phys. B **324**, 106 (1989).

On the linearization stability of the conformally (anti-) self-dual Einstein equations

C. G. Torre[a)]
Center for Space Science Research, Space Research Institute, Florida Institute of Technology, Melbourne, Florida 32901-2715

(Received 26 April 1990; accepted for publication 8 August 1990)

The Einstein equations with a cosmological constant, when restricted to Euclidean space-times with anti-self-dual Weyl tensor, can be replaced by a quadratic condition on the curvature of an SU(2) (spin) connection. As has been shown elsewhere, when the cosmological constant is positive and the space-time is compact, the moduli space of gauge-inequivalent solutions to this equation is discrete, i.e., zero dimensional; when the cosmological constant is negative, the dimension of the moduli space is essentially controlled by the Atiyah–Singer index theorem provided the field equations are linearization stable. It is shown that linearization instability occurs whenever the unperturbed geometry possesses a Killing vector and/or a "harmonic Weyl spinor." It is then proven that while there are no Killing vectors on compact conformally anti-self-dual Einstein spaces with a negative cosmological constant, it *is* possible to have harmonic Weyl spinors. Therefore, the conformally anti-self-dual Einstein equations on a compact Euclidean manifold are linearization stable when the cosmological constant is negative provided the unperturbed geometry admits no harmonic Weyl spinors.

I. INTRODUCTION

Recent work of Samuel[1] as well as Capovilla, Dell, and Jacobson[2] has shown that the Euclidean signature Einstein equations with a nonzero cosmological constant, when restricted to geometries with an anti-self-dual Weyl tensor, can be replaced by five quadratic conditions on the curvature of an SU(2) spin connection on the space-time manifold M:

$$\tfrac{1}{4} F_{[ab}{}^{(AB} F_{cd]}{}^{CD)} = 0. \tag{1.1}$$

In the equation above, F_{ab} is the curvature of the left-handed spin connection; lower-case Latin indices are abstract space-time indices while capital Latin indices are abstract SU(2) spinor indices that are raised and lowered with the antisymmetric spinor ϵ^{AB} and its inverse. As shown in Ref. 2, (1.1) is equivalent to the statement that

$$F_{ab}{}^{AB} = -\tfrac{1}{6}\lambda \Sigma_{ab}{}^{AB}, \tag{1.2}$$

where

$$\Sigma_{ab}{}^{AB} := 2\gamma_{[a}{}^{AA'}\gamma_{b]A'}{}^{B}, \tag{1.3}$$

and λ is the cosmological constant.[3] In (1.3) γ_a is an SU(2)×SU(2) soldering form which defines the metric[4] via

$$g_{ab} = \gamma_a{}^{AA'}\gamma_{bAA'}; \tag{1.4}$$

the metric in turn defines a Hodge duality operation with respect to which Σ_{ab}—and, from (1.2), F_{ab}—are self-dual. In a solution to (1.1), the curvature F_{ab} corresponds to the self-dual part of the Riemann tensor; (1.2) then implies that the Weyl tensor is anti-self-dual. Conversely, every conformally anti-self-dual Einstein space arises as a solution to (1.1).[5]

In Ref. 6 we began a study of the space of solutions to the conformally anti-self-dual Einstein equations by analyzing the linearized version of (1.1), which is given by

$$D_1 C := F_{[ab}{}^{(AB} D_c C_{d]}{}^{CD)} = 0. \tag{1.5}$$

Here, C_a is the perturbation of the left-handed spin connection and D_a is the corresponding (unperturbed) derivative operator with curvature F_{ab}; Eq. (1.5) is obtained only if the unperturbed curvature satisfies (1.1). The linearized equation (1.5) admits an infinite number of solutions that are generated by the action of the "gauge group" of general relativity which, in the formalism being used here, is a semidirect product of the diffeomorphism and local SU(2) groups. Infinitesimal gauge transformations correspond to perturbations of the form (see Ref. 6 for details)

$$C_a = D_0(f,M,N) := (\nabla^b f)F_{ba} + [D^b M, F_{ba}] + D_a N, \tag{1.6}$$

where N and M are su(2)-valued functions[7] and f is a real-valued function. Here, and in what follows, space-time indices are lowered and raised by the metric associated with the unperturbed solution to (1.1). Notice that we are using an su(2) matrix notation which suppresses spinor indices, e.g., the bracket in (1.6) is an su(2) commutator. As all perturbations of the form (1.6) solve (1.5), one is naturally lead to study gauge-inequivalent solutions of the linearized equations; they are the equivalence classes

$$[C] = \text{kernel } D_1/\text{image } D_0. \tag{1.7}$$

One of the central results of Ref. 6 was that these equivalence classes arise as the kernel of an elliptic operator D:

$$[C] = \text{kernel } D, \tag{1.8}$$

where

$$D = (D_1, D_0^*), \tag{1.9}$$

$$D_0^* C = (\text{tr } F^{ab} D_a C_b; [D_a C_b, F^{ab}]; -D^a C_a). \tag{1.10}$$

The operator in Eq. (1.10) corresponds to the L^2 adjoint of D_0; here and in what follows we extend the action of D_a to

[a)] Current address: Physics Dept., Syracuse Univ., Syracuse, NY 13244-1130.

 0022-2488/90/122983-04$03.00

include tensor indices via the unique torsion-free derivative operator compatible with the metric obtained from the solution to (1.1). Because D is an elliptic operator on a compact manifold without boundary (we shall only work with such manifolds) it is clear that the gauge-inequivalent solutions to (1.5) form a finite-dimensional subspace of all possible perturbations. Furthermore, it was shown[6] that when the cosmological constant is positive (1.5) admits only trivial (pure gauge) solutions; thus the moduli space of left-handed spin connections on conformally anti-self-dual Einstein spaces is discrete, i.e., zero dimensional, in this case. It was pointed out that the dimension of $[C]$ in the $\lambda<0$ case could be determined via the Atiyah–Singer index theorem, however, the utility of this result for a determination of the dimension of the corresponding moduli space of gauge-inequivalent solutions to (1.1) depends on the linearization stability of (1.1), i.e., on whether every solution of (1.5) is an approximation to a solution of (1.1) modulo gauge transformations. As shown in Ref. 6, Eq. (1.1) is linearization stable if the adjoint of the operator D has a trivial kernel. Here, the adjoint is defined as

$$D^* = (D_1^*, D_0), \tag{1.11}$$

where D_1^*, which is the L^2 adjoint of D_1, acts on totally symmetric, valence-four spinor-valued four-forms ω via

$$D_1^*\omega = F^{cd}{}_{CD} D^b \omega_{abcd}{}^{ABCD}. \tag{1.12}$$

We have used the unperturbed equations, in particular (1.2), to simplify (1.12).

Our purpose in this paper is to analyze the issue of linearization stability by studying

$$\text{kernel } D^* = \text{kernel } D_0 \cup \text{kernel } D_1^*; \tag{1.13}$$

Eq. (1.1) is linearization stable if (1.13) is trivial. The existence of a kernel for D_0 corresponds to the existence of infinitesimal automorphisms of the SU(2) bundle over M that leave the connection invariant, i.e., elements of kernel D_0 are gauge symmetries of the spin connection. The correlation of linearization stability and the absence of symmetries is familiar from the Cauchy problem in (Lorentzian) general relativity.[8] However, unlike the situation which arises with the full Einstein equations, linearization stability of the restriction to the self-dual sector also involves the kernel of $D^*{}_1$, which turns out to correspond to "harmonic Weyl spinors" on M. We will treat each of these cases in the following two sections.

II. INFINITESIMAL GAUGE SYMMETRIES

Infinitesimal gauge symmetries, corresponding to elements of kernel D_0, can also be identified with the kernel of the elliptic operator[6]

$$\Delta_0 := D_0^* D_0, \tag{2.1}$$

which leads to a coupled set of differential equations for f, M, and N. The results we desire can be obtained more easily if we follow a somewhat more indirect route, so we begin by studying gauge transformations of the form

$$C_a = M^b F_{ba} + D_a N, \tag{2.2}$$

where M^a is the vector field corresponding to an infinitesimal diffeomorphism. As shown in Ref. 6, (1.6) is equivalent to (2.2) on conformally anti-self-dual Einstein spaces with a positive cosmological constant; when λ is negative (1.6) fails to be equivalent to (2.2) if M_a has a harmonic contribution to its Hodge decomposition. We will see below that this disparity is irrelevant for the characterization of infinitesimal symmetries.

Using (1.2) in (2.2), and then solving

$$-\tfrac{1}{6}\lambda M^b \Sigma_{ba} + D_a N = 0 \tag{2.3}$$

for M_a, we find a necessary condition for the existence of an infinitesimal symmetry to be

$$M_a = -(1/\lambda)\text{tr}\,\Sigma_a{}^b D_b N = -(1/\lambda)\nabla_b(\text{tr}\,\Sigma_a{}^b N), \tag{2.4}$$

which, in particular, implies that M_a is coexact. From (2.4) we see that if M^a generates the diffeomorphism part of a gauge symmetry it can have no harmonic contribution to its Hodge decomposition, hence the set of all solutions to (2.3) is equivalent to kernel D_0. Furthermore, (2.4) implies that the term involving the function f in (1.6) must vanish since this term comes from the exact part of M_a (Ref. 6).

Substituting (2.4) into (2.3) leads to a necessary and sufficient condition for N to generate the local SU(2) part of a gauge symmetry:

$$d_0 N := D_a N + \tfrac{1}{4}[\Sigma_a{}^b, D_b N] = 0. \tag{2.5}$$

The L^2 adjoint of d_0 is given by

$$d_0^* C = -D^a C_a + \tfrac{1}{4}[D_a C_b, \Sigma^{ab}], \tag{2.6}$$

and we have

$$N \in \text{kernel } d_0 = \text{kernel } d_0^* d_0 \Leftrightarrow -D^a D_a N - \tfrac{2}{3}\lambda N = 0. \tag{2.7}$$

From (2.7) we see that the operator $d_0^* d_0$ is positive definite when the cosmological constant is negative, thus in this case the only solution to (2.7) is $N=0$; this forces M^a to vanish also, so there are no infinitesimal symmetries of the spin connection on compact conformally anti-self-dual Einstein spaces with a negative cosmological constant.

We can obtain a more transparent geometrical interpretation of this result if we look for the necessary and sufficient restriction on the vector field M^a to yield an infinitesimal symmetry. Beginning again with (2.3), we now view it as an equation to be solved for N with M^a treated as given; we find

$$N = \tfrac{1}{4}\Sigma^{ab}\nabla_a M_b, \tag{2.8}$$

which implies that N is the spinor representation of the self-dual part of the exterior derivative of M_a. After substituting (2.8) into (2.3) we obtain

$$-\nabla^b\nabla_b M^a + \nabla^a\nabla_b M^b - \lambda M^a = 0. \tag{2.9}$$

By taking the divergence of both sides of (2.9) we deduce

$$\nabla_a M^a = 0, \tag{2.10}$$

which is consistent with (2.4). In the presence of (2.10), Eq. (2.9) becomes

$$-\nabla^b\nabla_b M^a - \lambda M^a = 0, \tag{2.11}$$

which, as before, has no solutions if $\lambda<0$. It is easily verified that (2.10) and (2.11) imply M^a satisfies the Killing equations on the compact Einstein space (M,g).

The results we have obtained in this section are analogous to the situation arising with compact two-dimensional (Riemann) surfaces. All such surfaces are Einstein spaces; genus-0, the sphere, has a positive cosmological constant, genus-1, the torus, has $\lambda = 0$, while all higher genus surfaces have a negative cosmological constant. The relevant symmetry group here is the group of conformal isometries; it is well-known that the sphere admits conformal isometries connected to the identity and, in addition, the moduli space of the sphere is trivial. For surfaces of genus greater than one, there are no infinitesimal conformal isometries, i.e., any conformal isometries are not in the connected component of the identity, while the moduli space for these surfaces is nontrivial.

III. HARMONIC WEYL SPINORS

We now turn to a study of the kernel of D_1^*, i.e., we analyze the solution space of

$$D_1^*\omega = F^{cd}{}_{CD} D^b \omega^{ABCD}_{abcd} = 0. \tag{3.1}$$

Keeping in mind that the unperturbed SU(2) curvature satisfies (1.2), Eq. (3.1) is equivalent to

$$D^b \omega_{ab} = 0, \tag{3.2}$$

where the su(2)-valued two-form ω_{ab} is defined by

$$\omega^{AB}_{ab} := \Sigma_{abCD} \epsilon^{lmnp} \omega^{ABCD}_{lmnp}. \tag{3.3}$$

From (3.3) it is clear that ω_{ab} is self-dual, thus (3.2) can be reexpressed as

$$D_{[a}\omega_{bc]} = 0. \tag{3.4}$$

Notice that Eqs. (3.2) and (3.4) are precisely the Yang–Mills equations (in the self-dual sector) for SU(2) gauge theory. Similar equations also arise in the linearization stability analysis of the self-duality condition in Yang–Mills theory,[9] however, there are two important flaws in the analogy between the gravitational and gauge theoretic treatments: (1) ω_{ab} is not quite an SU(2) curvature—it has the symmetries of the (self-dual) Weyl spinor; in particular, ω_{ab} has only five independent components while a self-dual SU(2) curvature has nine independent components; (2) the linearization stability analysis of the self-dual Yang–Mills equations actually yields Eqs. (3.2) and (3.4) as conditions on an *anti*-self-dual su(2)-valued two-form. Still, we shall now show that these two differences, so to speak, cancel each other, and we arrive at a vanishing theorem quite analogous to that arising in Yang–Mills theory.

We proceed using what is by now a familiar strategy. Denote by d_1^* the differential operator appearing in (3.2):

$$d_1^*\omega := 2D^b \omega_{ab}. \tag{3.5}$$

The L^2 adjoint of d_1^* is given by

$$d_1 C = (\delta^{[a}_c \delta^{b]}_d + \tfrac{1}{2}\epsilon_{cd}{}^{ab}) D_a C_b, \tag{3.6}$$

and we have

$$\text{kernel } d_1^* = \text{kernel } d_1 d_1^*. \tag{3.7}$$

Explicit computation, which makes use of the instanton equation (1.2) and the definition (3.3), reveals

$$\omega_{ab} \in \text{kernel } d_1 d_1^* \Leftrightarrow -D^c D_c \omega_{ab} + 2\lambda \omega_{ab} = 0. \tag{3.8}$$

Equation (3.8), when rewritten in terms of the totally symmetric spinor

$$\omega_{ABCD} := \epsilon^{abcd} \omega^{ABCD}_{abcd}, \tag{3.9}$$

is equivalent to

$$-D^c D_c \omega_{ABCD} + 2\lambda \omega_{ABCD} = 0; \tag{3.10}$$

thus Eqs. (3.2) and (3.3) are satisfied when the manifold admits harmonic Weyl spinors.[10] The terminology "harmonic Weyl spinor" is meant to be suggestive of (3.10) but is not to be taken too literally: *the* self-dual Weyl spinor, which appears in the spinor decomposition of the Riemann tensor, is required to vanish on the space-times we are studying here. We call ω_{ABCD} a Weyl spinor only because it possesses all the algebraic symmetries of the Weyl (conformal) curvature spinor. Similarly, the term "harmonic" is not to be interpreted in the usual sense of Hodge–de Rham theory, but simply implies that ω_{ABCD} satisfies the most natural elliptic differential equation compatible with its algebraic symmetries.

When the cosmological constant is positive, the operator on the left-hand side of (3.10) is positive definite, thus there are no harmonic (self-dual) Weyl spinors on conformally anti-self-dual Einstein spaces with a positive cosmological constant. Unfortunately, there is no general obstruction to solutions of (3.10) when $\lambda < 0$, i.e., linearization stability is not guaranteed for $\lambda < 0$.

IV. LINEARIZATION STABILITY

For the convenience of the reader we will now assemble the results of the preceding sections. It is natural to classify the results by the sign of the cosmological constant.

A. $\lambda > 0$

Linearization stability is not really an issue here because there are no nontrivial solutions to the linearized equations. Nevertheless, we have found that $\lambda > 0$ is compatible with the existence of gauge symmetries of the connection; these correspond to the existence of Killing vectors on M. In addition, there are no harmonic Weyl spinors in this case.

B. $\lambda < 0$

In this case the situation is the reverse of the previous results. There are no infinitesimal symmetries of the connection; this corresponds to the absence of Killing vectors on compact Einstein spaces with a negative cosmological constant. On the other hand, there is no general obstruction to the existence of harmonic Weyl spinors when $\lambda < 0$; we conclude that the conformally anti-self-dual Einstein equations are linearization stable whenever there exist no Weyl spinors obeying the (eigenvalue) condition (3.10).

V. DISCUSSION

The linearization stability analysis of the conformally anti-self-dual Einstein equations is remarkably similar to the situation occurring in SU(2) Yang–Mills theory.[9] The gravitational results *do*, however, differ from those of gauge theory owing to the presence of the diffeomorphism group as a symmetry group as well as the (related) fact that the gravitational SU(2) curvature perturbation is required to be *anti-*

self-dual even though the unperturbed SU(2) curvature is self-dual.

As pointed out in Ref. 6, the relationship between the space of gauge symmetries, gauge inequivalent perturbations, and harmonic Weyl spinors is controlled by the Atiyah–Singer index theorem. If we denote the topological index as I, then we now know that

$$I = \dim \text{kernel } D_0; \quad \lambda > 0, \tag{5.1a}$$

$$I = \dim \text{kernel } D_1^* - \dim(\text{kernel } D_1 \cap \text{kernel } D_0^*); \lambda < 0, \tag{5.1b}$$

where the second term in (5.1b) is the dimension of the space of gauge-inequivalent perturbations satisfying (1.5). Clearly, it will be necessary to investigate the question of how "generic" is the existence of Weyl spinors satisfying the eigenvalue condition (3.10). The type of result which one might be able to obtain could be analogous to the fact that "If M is a connected, compact orientable Einstein space of scalar curvature 1, then M admits an eigenfunction f with $\Delta f = -nf$ if and only if M is isometric to $S^n(1)$ [the n-sphere of unit radius]".[11] In any event, it is clear that (5.1) may provide more than just a way to calculate the dimension of the moduli space of left-handed spin connections; indeed, the sign of I can represent a topological obstruction to the existence of a conformally anti-self-dual Einstein metric. For example, if the topology of M is such that $I < 0$, then (5.1) cannot be satisfied for any such metric with $\lambda > 0$. Alternatively, if the space-time admits no solutions to (3.10) and $I > 0$, then (5.1) cannot be satisfied when $\lambda < 0$. Another interesting corollary to (5.1) is that the topological index controls the dimension (of the Lie algebra) of the isometry group when $\lambda > 0$. The topological implications of (5.1) will be explored in a future publication.

ACKNOWLEDGMENT

This work was supported in part by a grant from the Florida Technological Research and Development Authority to the Florida Institute of Technology.

[1] J. Samuel, Class. Quantum Gravit. **5**, L123 (1988).
[2] R. Capovilla, J. Dell, and T. Jacobson, Class. Quantum Gravit. **7**, L1 (1990).
[3] We have chosen our conventions such that the cosmological constant is -6 times that of Ref. 1. Our conventions are consistent with the Einstein equations $R_{ab} = \lambda g_{ab}$.
[4] Because (1.1) is independent of the metric, it is possible to have solutions corresponding to a degenerate metric. Throughout this paper we only consider solutions of (1.1) which correspond to nondegenerate metrics.
[5] S. Koshti and N. Dadhich, Class. Quantum Gravit. **7**, L5 (1990). Identical results for conformally self-dual space-times can be obtained by working with the right-handed spin connection.
[6] C. G. Torre, Phys. Rev. D. **41**, 3620 (1990).
[7] Notice that we notationally distinguish the group SU(2) from its Lie algebra su(2).
[8] V. Moncrief, J. Math. Phys. **16**, 493 (1975).
[9] M. Atiyah, N. Hitchin, and I. Singer, Proc. R. Soc. London Ser. A **362**, 425 (1978); A. Schwarz, Commun. Math. Phys. **64**, 233 (1979).
[10] We could, of course, have arrived at (3.10) directly, i.e., without the use of the su(2)-valued two-form. The introduction of this two-form was intended to strengthen the analogy with the results of the self-duality analysis of SU(2) gauge theory.
[11] S. Kobayashi and K. Nomizu, *Foundations of Differential Geometry* (Interscience, New York, 1969), Vol. II.

On the structure of quantum phase space

R. Aldrovandi and D. Galetti
Instituto de Física Teórica, Universidade Estadual Paulista, Rua Pamplona, 145, 01405-São Paulo, SP, Brazil

(Received 4 January 1990; accepted for publication 11 July 1990)

The space of labels characterizing the elements of Schwinger's basis for unitary quantum operators is endowed with a structure of symplectic type. This structure is embodied in a certain algebraic cocycle, whose main features are inherited by the symplectic form of classical phase space. In consequence, the label space may be taken as the quantum phase space: It plays, in the quantum case, the same role played by phase space in classical mechanics, some differences coming inevitably from its nonlinear character.

I. INTRODUCTION

The recent extension of Weyl–Wigner transformations to discrete quantum spectra[1] has drawn attention to a certain discrete space with some characteristics of a "quantum phase space" (QPS).[2] The extension makes use of Schwinger's complete basis[3] of unitary operators for Weyl's realization of the Heisenberg group. Unlike usual classical phase spaces, QPS is not a linear space: Its points, besides being isolated, display themselves on the surface of a torus. The continuum quantum case may be obtained by a standard procedure that corresponds to stretching the torus radii to infinity while bringing the spacing between neighboring points to zero in a suitable way. This C-number representation of QPS closely parallels the classical picture, its quantum character being signaled by the presence of Planck's constant $\hbar$ in the expressions involved. It is of basic interest to examine the main properties of QPS and their relations to the well-known characteristics of the classical phase space. We would of course expect to obtain the classical case as a $\hbar \to 0$ limit of the quantum case.

The basic feature of a classical phase space is its symplectic structure, embodied in a differential two-form Ω which is closed (a cocycle) and nondegenerate. The fundamental role of this symplectic form is especially visible in the Hamiltonian formulation of mechanics. So strongly does the symplectic structure stick to the very notion of phase space that QPS will only deserve its name if it includes a structure of similar nature. Although we may not expect the presence of a complete analog to Ω on QPS, our objective here is to show that a certain structure exists indeed which plays on QPS a role as similar to a symplectic structure as could be expected. Such a "presympletic" structure is actualized in a certain two-cochain (also a cocycle) acting on the unitary operators, a purely algebraic object which acquires, in the continuous limit, a geometrical nature and tends, in the classical limit, to the symplectic form. The two-cochain marks in reality the projective character of Weyl's realization of the Heisenberg group.

We start in Sec. II with a sketchy presentation of Hamiltonian mechanics[4] intended to fix notation for later comparison, special emphasis being given to the role of the symplectic structure.[5] We then address ourselves to quantum kinematics and give a resumé on Schwinger's complete basis of unitary operators in Sec. III. A crucial point will be that the basis provides in reality not a linear but a projective representation of the Heisenberg group. Preparing to establish that, Sec. IV is a short introduction to the subject of projective representations[6] from the cohomological point of view[7] which, being closer to the formalism of differential forms, is specially convenient to our purposes.[8] The meaning of ray representations becomes specially clear in this language. The results are then applied in Sec. V to the Schwinger basis for the Weyl representation, emphasis being given to the emergence of the mentioned cocycle and to some of its properties. The continuum limit is examined and comparison is made with another C-number representation of quantum mechanism, the Weyl–Wigner–Moyal[9] approach. The meaning of the "presymplectic" fundamental cocycle is clarified in terms of well-known features of that approach.

II. CLASSICAL PHASE SPACE

In the classical description of a system with n degrees of freedom, physical states constitute a differentiable symplectic manifold M of dimension $2n$. The fundamental geometrical characteristic of this phase space is the *symplectic two-form* Ω. In terms of the generalized coordinates $q = (q^1, q^2, \ldots, q^n)$ and momenta $p = (p_1, p_2, \ldots, p_n)$, Ω is written

$$\Omega = dq^i \wedge dp_i. \tag{2.1}$$

It is clearly a closed form (that is, $d\Omega = 0$), or *cocycle*, and can be shown also to be nondegenerate. Here, Ω is also an exact form (a *coboundary*, or a trivial cocycle) as it is, up to a sign, the differential of the *canonical* form

$$\sigma = p_i\, dq^i. \tag{2.2}$$

The structure defined by a closed nondegenerate two-form is called a *symplectic structure* and a manifold endowed with such a structure is a *symplectic manifold*. In reality, phase spaces are very particular cases of symplectic manifolds. On general, topologically nontrivial symplectic manifolds there are no global coordinates such as the (q^i, p_i) supposed above and the basic closed nondegenerate two-form is not necessarily exact. Notice that every coboundary is a cocycle but not vice versa. A theorem by Darboux

 0022-2488/90/122987-09$03.00

ensures the existence of a chart (of "canonical," or "symplectic" coordinates) around any point on a $(2n)$-dimensional manifold M in which a closed nondegenerate two-form can be written as in (2.1), so that the equations here written in components hold locally. Notice, however, that Ω is globally defined and the equations written in the invariant language of forms are valid globally.

The fundamental point about the symplectic structure is that Ω establishes a one-to-one relationship between one-forms and vector fields on the manifold M. The simplest example is the phase space velocity field,

$$X_H=\frac{dq^i}{dt}\frac{\partial}{\partial q^i}+\frac{dp_i}{dt}\frac{\partial}{\partial p_i}. \tag{2.3}$$

The time evolution of the state point (q,p) will take place along the integral curves of X_H. Hamilton's equations put this evolution field into the form

$$X_H=\frac{\partial H}{\partial p_i}\frac{\partial}{\partial q^i}-\frac{\partial H}{\partial q^i}\frac{\partial}{\partial p_i}. \tag{2.4}$$

The differential operator X_H generates a one-parameter group of transformations, the *Hamiltonian flow*. On the other hand, the Hamiltonian function $H(q,p)$ will have as differential the one-form

$$dH=\frac{\partial H}{\partial p_i}dp_i+\frac{\partial H}{\partial q^i}dq^i. \tag{2.5}$$

The relationship for which Ω is responsible involves the *interior product* of a field by a form. The interior product of a field X by a one-form σ, denoted $i_X\sigma$, is simply $\sigma(X)$. The interior product of a field X by a two-form Ω, denoted $i_X\Omega$, is defined as that one-form satisfying $i_X\Omega(Y)=\Omega(X,Y)$ for any field Y. This is directly generalized to higher-order forms. We find easily that

$$i_{X_H}\Omega=dH. \tag{2.6}$$

Besides being a particular case of the general one-to-one relationship between fields (vectors) and one-forms (covectors) on M, this is also an example of relationship between a transformation generator and the corresponding generating function. The Hamiltonian presides over the time evolution of the physical system under consideration: $H(q,p)$ is the *generating function* of the velocity field X_H. Applying X_H to any given differentiable function $F(q,p)$ on M, we find that

$$X_HF=\frac{\partial F}{\partial q^i}\frac{\partial H}{\partial p_i}-\frac{\partial F}{\partial p_i}\frac{\partial H}{\partial q^i}=\{F,H\}, \tag{2.7}$$

the Poisson bracket of F and H, so that its equation of motion is the Liouville equation

$$\frac{dF}{dt}=X_HF, \tag{2.8}$$

X_H is frequently called *Liouvillian operator*. Functions like $F(q,p)$ are the classical observables, or dynamical functions. To each such a function will correspond a field

$$X_F=\frac{\partial F}{\partial p_i}\frac{\partial}{\partial q^i}-\frac{\partial F}{\partial q^i}\frac{\partial}{\partial p_i} \tag{2.9}$$

through the relation

$$i_{X_F}\Omega=dF. \tag{2.10}$$

Given another function $G(q,p)$ and its corresponding field X_G, it is immediate to verify that

$$\Omega(X_F,X_G)=\frac{\partial F}{\partial q^i}\frac{\partial G}{\partial p_i}-\frac{\partial F}{\partial p_i}\frac{\partial G}{\partial q^i}=\{F,G\}. \tag{2.11}$$

Each field on M is the local generator of a one-dimensional group of transformations. The response of a tensor to the local transformations generated by a field is measured by the Lie derivative of the tensor with respect to the field. Of course, F (which is a zero-order tensor) is an integral of motion if its Lie derivative $\mathbf{L}_{X_H}F=X_HF$ vanishes, or $\{F,H\}=0$. The Lie derivative of Ω with respect to X_H vanishes:

$$\mathbf{L}_{X_H}\Omega=0, \tag{2.12}$$

because $\mathbf{L}_X=d\circ i_X+i_X\circ d$. This means that the two-form Ω is preserved by the Hamiltonian flow, or by the time evolution. This and the property $\mathbf{L}_X(\Omega\wedge\Omega)=(\mathbf{L}_X\Omega)\wedge\Omega+\Omega\wedge(\mathbf{L}_X\Omega)$ of Lie derivatives establish the invariance of the whole series of Poincaré invariants $\Omega\wedge\Omega\cdots\wedge\Omega$, including that with a number of n of Ω's, which is proportional to the volume form of M. The preservation of the volume form by the Hamiltonian flow is of course Liouville's theorem.

For any field X_F related to a dynamical function F,

$$\mathbf{L}_{X_F}\Omega=0. \tag{2.13}$$

This happens because

$$\mathbf{L}_{X_F}\Omega=d\circ i_{X_F}\Omega+i_{X_F}\circ d\Omega=d^2F=0.$$

Such transformations leaving Ω invariant are the *canonical transformations*, X_F is said to be a *Hamiltonian field* and F its *generating function*. In a more usual language, F is the generating function of the corresponding canonical transformation. The simplest examples of generating functions are given by $F(q,p)=q^i$, corresponding to the field $X_F=-\partial/\partial p_i$; and $G(q,p)=p_i$, whose field is $X_G=\partial/\partial q^i$. Both lead to $\{q^i,p_j\}=\delta^i_j$. Next in simplicity are the dynamical functions of the type

$$f_{ab}=aq+bp, \tag{2.14}$$

with a, b real constants. The corresponding fields are $J_{ab}=-a\,\partial/\partial p+b\,\partial/\partial q$. The commutator of two such fields is $[J_{ab},J_{cd}]=0$ and consequently the corresponding generating function $F_{[J_{ab},J_{cd}]}=F_0$ is a constant. On the other hand, the Poisson brackets are determinants

$$\{f_{ab},f_{cd}\}=\Omega(J_{ab},J_{cd})=ad-bc. \tag{2.15}$$

With the fields written as

$$X=\begin{pmatrix}X_{q^i}\\X_{p_i}\end{pmatrix},$$

Ω and Ω^{-1} may be seen as matrices:

$$\Omega=\begin{bmatrix} 0 & \mathbf{I}_n \\ -\mathbf{I}_n & 0 \end{bmatrix}; \quad \Omega^{-1}=\begin{bmatrix} 0 & -\mathbf{I}_n \\ \mathbf{I}_n & 0 \end{bmatrix}, \tag{2.16}$$

where $\mathbf{I}_n$ is the n-dimensional unit matrix and $\Omega(X,Y) = X^T\Omega Y$.

Most fields do not correspond to a generating function, as $i_X\Omega$ is not always exact. In general, a generating function exists only locally. The one-form corresponding to any field preserving Ω will be closed, $d(i_X\Omega) = L_X\Omega = 0$. As a closed form is always locally exact, around any point of M there is a neighborhood where some $F(q,p)$ satisfies $i_X\Omega = dF$.

The action of the two-form Ω on two contravariant fields X and Y will give

$$\Omega(X,Y)=X_{q^i}Y_{p_i}-X_{p_i}Y_{q^i}. \tag{2.17}$$

This is twice the area of the triangle defined on M by X and Y, as it is still easier to see from (2.14) and (2.15).

An n-dimensional subspace of the $2n$-dimensional phase space M is a *Lagrange manifold* if $\Omega(X,Y) = 0$ for any two vectors X, Y tangent to it. Examples are the configuration space and the momentum space. Canonical transformations preserve such subspaces of M, that is, they take a Lagrange manifold into another Lagrange manifold.

The symplectic form being a cocycle is equivalent to the Jacobi identity for the Poisson brackets. In fact, it is not difficult to find that

$$\begin{aligned} 3\,d\Omega(X,Y,Z) = &-\{F_X,\{F_Y,F_Z\}\}-\{F_Z,\{F_X,F_Y\}\} \\ &-\{F_Y,\{F_Z,F_X\}\}=0. \end{aligned} \tag{2.18}$$

There would be of course much more to be said about phase space. This brief outline, however, seems enough to establish notation and stress the basic role of the cocycle Ω. We shall see in Sec. V that on quantum phase space a cocycle is also defined which, even in the discrete case, has a comparably fundamental role.

III. QUANTUM KINEMATICS

The quantum description of a physical system requires a complete set of observables. Still better, it requires a complete set of operators in terms of which all dynamical operators can be built up. Kinematics is governed by Heisenberg's group,[10] whose elements may be represented by real triples (a,b,r) obeying the group product rule[11]

$$(a,b,r)*(c,d,s)=(a+c,b+d,r+s+\tfrac{1}{2}[ad-bc]).$$

Weyl introduced a realization in terms of powers of two unitary operators $U(a)$ and $V(b)$ satisfying

$$U(a)U(a')=U(a+a'),$$

$$V(b)V(b')=V(b+b'),$$

and

$$U(a)V(b)=V(b)U(a)e^{i\mu ab}.$$

A particular example is given by $V = e^{ib\mathbf{p}}$, $U = e^{ia\mathbf{q}}$, which lead to the usual formulation of Heisenberg's algebra using the basic operators $\mathbf{p}$ and $\mathbf{q}$. Schwinger[3] has recognized the fact that the above U and V generate a complete basis for all unitary operators and provided a classification of all the possible physical degrees of freedom. We shall here be interested only in some aspects of Schwinger's work. What follows is a short presentation of them.

Consider a space of quantum states of which a basis is given by orthonormalized kets $|v_k\rangle$ with $k = 1, 2,..., N$. A unitary operator U can be defined which shifts these kets through cyclic permutations as

$$U|v_k\rangle=|v_{k+1}\rangle, \quad \text{with } |v_{k+N}\rangle\equiv|v_k\rangle. \tag{3.1}$$

Through the repeated action of U, a set of linearly independent unitary operators U^m can be obtained whose action is given by

$$U^m|v_k\rangle=|v_{k+m}\rangle. \tag{3.2}$$

As $U^N = \mathbf{1}$, the eigenvalues of U are $u_k = e^{i(2\pi/N)k}$, corresponding to another set of kets fixed by

$$U|u_k\rangle=u_k|u_k\rangle. \tag{3.3}$$

Another operator V exists such that

$$V|u_k\rangle=|u_{k-1}\rangle \tag{3.4}$$

and

$$V^n|u_k\rangle=|u_{k-n}\rangle, \quad \text{with}\,|u_{k-N}\rangle\equiv|u_k\rangle. \tag{3.5}$$

Here, also, $V^N = \mathbf{1}$ and the V eigenvalues are $v_k = e^{i(2\pi/N)k}$. The miracle of Schwinger's basis is that the eigenkets $|v_k\rangle$ such that

$$V|v_k\rangle=e^{i(2\pi/N)k}|v_k\rangle \tag{3.6}$$

are just those from which we have started. Of course,

$$V^n|v_k\rangle=e^{i(2\pi/N)kn}|v_k\rangle. \tag{3.7}$$

A direct calculation in any basis shows that

$$V^nU^m=e^{i(2\pi/N)mn}U^mV^n. \tag{3.8}$$

Now, Schwinger's final point: The set of operators

$$S_{mn}=e^{i(\pi/N)mn}U^mV^n \tag{3.9}$$

constitute a complete orthogonal basis in terms of which any dynamical quantity $\mathbf{O}$ can be constructed as

$$\mathbf{O}=\sum_{m,n} O_{mn}S_{mn}, \tag{3.10}$$

the O_{mn}'s being coefficients given by

$$O_{mn}=\mathrm{tr}[S_{mn}^{\dagger}\mathbf{O}]. \tag{3.11}$$

Here, U and V are each one a generator of the cyclic group Z_N. The operators S_{mn} give a peculiar combination of the two Z_N's, providing a discrete version of Weyl's representation of the Heisenberg group.

The following results are easily obtained: (i) the action of the basic operators on the kets:

$$S_{mn}|v_k\rangle=e^{i(\pi/N)(2k+m)n}|v_{k+m}\rangle; \tag{3.12}$$

(ii) the group product:

$$S_{rs}S_{mn}=e^{i(\pi/N)(ms-nr)}S_{(m+r)(n+s)}; \tag{3.13}$$

(iii) the group identity:

$$S_{00}\equiv\mathbf{1}; \tag{3.14}$$

(iv) the inverse to a given element:

$$S^{-1}{}_{mn}=S_{-m,-n}; \tag{3.15}$$

(v) behavior under a similarity transformation:

$$S_{mn}S_{rs}S^{-1}{}_{mn}=e^{-i(2\pi/N)(ms-nr)}S_{rs}. \tag{3.16}$$

(vi) associativity:

$$(S_{mn}S_{rs})S_{kl}=S_{mn}(S_{rs}S_{kl}). \tag{3.17}$$

With the periodicity conditions in (3.1) and (3.5), the numbers m, n, etc. take values on a torus lattice. It is this lattice who plays the role of a quantum phase space. The points of QPS are so labels of elements of a discrete group. The operators S_{mn}, obeying the product rule (3.13), give a projective representation of the group of transformations on this space, which will be examined in the next section. Notice that they are themselves only semiperiodical: $S_{Np}=(-)^pS_{0p}$; $S_{pN}=(-)^pS_{p0}$. The quantum continuum limit, which has only been studied in detail in some cases,[1,3] is in such cases attained by taking both the torus radii to be infinite while making the spacing between neighboring points go to zero, in such a way that $[\sqrt{2\pi/N}m]\to$some real constant a, $[\sqrt{2\pi/N}n]\to$another real constant b, etc. In this limit, a particular realization of the above operators is

$$V=e^{i\sqrt{2\pi/N}\mathbf{p}}; \quad U=e^{i\sqrt{2\pi/N}\mathbf{q}}, \tag{3.18}$$

$$V^n\to e^{ib\mathbf{p}}; \quad U^m\to e^{ia\mathbf{q}}, \tag{3.19}$$

where the operators $\mathbf{p}$ and $\mathbf{q}$ have eigenvalues $\sqrt{2\pi/N}k$. In this case,

$$S_{mn}\to S_{ab}=e^{i(a\mathbf{q}+b\mathbf{p})}. \tag{3.20}$$

The expression (3.13) takes the form

$$S_{cd}S_{ab}=e^{(i/2)[ad-cb]}S_{(a+c)(b+d)}. \tag{3.21}$$

The exponent in S_{ab} is the quantum version of the dynamical functions (2.14) and the phase in the group product is just (half) the Poisson bracket (2.15).

To a given degree of freedom corresponds a pair of operators U, V satisfying (3.8) which will provide a basis for a realization of the Heisenberg group. A curious and important example is given by the nonlocal order and disorder operators determining the confined and unconfined phases in quarkionic matter.[12] The algebra (3.8) appears then because of the crucial role attributed to the center of the group SU(N), which is precisely Z_N.

The above considerations on the continuum limit suggest that each pair of operators U, V satisfying (3.8) is related to a pair of (exponentiated) canonically conjugate variables and, so, to a degree of freedom. This is true only when N is a prime number.[3] Otherwise, the representation involved is reducible. When N is prime, (3.8) is the only possible combination of powers of U and V leading to such a kind of expression. When N is not prime, however, things are different; N can be written in terms of its prime factors, $N=N_1N_2\cdots N_j$ and particular powers of U and V combine to give expressions like (3.8) with N replaced by each one of the factors N_i. The basis can then be redefined to become a direct product.[3] In the continuous limit, N goes to infinity through prime values.

IV. PROJECTIVE REPRESENTATIONS

Projective representations[6] are treated, even in the best of older physicists' texts, in a rather involved way. The modern, homological approach[7] of which a brief account is given in the following has many advantages, not the least being its assignment of the subjet's correct place in the wider chapter of group extensions.[8] In our case the main advantage is that the evident analogy with the formalism of differential forms allows a clearer view of the connections between Schwinger's basis and classical phase space.

Let us consider, to fix the ideas, a group G of elements g, h, etc., acting through their representative operators $U(g)$, $U(h)$, etc. on kets $|\varphi_x\rangle$, $|\psi_y\rangle$, etc. The indices x, y include not only configuration or momentum space coordinates but also spin and/or isospin indices and any other necessary state labels. We shall call them *parameters*. We might alternatively talk of the corresponding wave functions $\varphi(r)=\langle r|\varphi\rangle$, etc., but will use kets to keep in pace with previous notation. The space $\{|\varphi_x\rangle\}$ of kets will be the carrier space of the representation.

Suppose to begin with that we have

$$U(g)|\varphi_x\rangle=|\varphi_{xg}\rangle, \tag{4.1}$$

where "xg" is the set of labels as transformed by the action of g. Suppose further that, by composition,

$$U(h)U(g)|\varphi_x\rangle=U(gh)|\varphi_x\rangle, \tag{4.2}$$

meaning, in particular, that the composition by itself is independent of the point x in parameter space. This is what is usually called a representation, but will in the present context be called a *linear representation*. The mapping $U:g\to U(g)$ is in this case a homomorphism.

We may next suppose that, instead of (4.1), the action of a transformation is given by

$$U(g)|\psi_x\rangle=e^{i\alpha_1(x;g)}|\psi_{xg}\rangle. \tag{4.3}$$

The wave function acquires a phase $\alpha_1(x;g)$ which depends both on the transformation and the point in parameter space. The transformation will operate differently at different x. In quantum mechanics, of course, a state is fixed by a *ray* (a wave function with any phase factor). A representation acting according to (4.3) has been called a *ray representation*. It is a particular case of *projective representation*, as will be seen in the following.

Suppose condition (4.2) holds,

$$U(h)U(g)|\psi_x\rangle=U(gh)|\psi_x\rangle.$$

A direct calculation shows that this implies

$$\alpha_1(xg;h)-\alpha_1(x;gh)+\alpha_1(x;g)=0, \tag{4.4}$$

another form of the homomorphic condition. If a function $\alpha_0(x)$ exists such that $\alpha_1(x;g)$ can be written in the form

$$\alpha_1(x;g)=\alpha_0(xg)-\alpha_0(x), \tag{4.5}$$

then (4.4) holds automatically, (4.3) becomes

$$U(g)e^{i\alpha_0(x)}|\psi_x\rangle=\mathbf{e}^{i\alpha_0(xg)}|\psi_{xg}\rangle$$

and phases can be eliminated by redefining

$$|\varphi_x\rangle=\mathbf{e}^{i\alpha_0(x)}|\psi_x\rangle,$$

which brings the group action back to the form (4.1).

In the cohomological theory of group representations, phases such as the above $\alpha_0(x)$ and $\alpha_1(x;g)$ are considered as results of the action of cochains on the group G. Cochains are antisymmetric mappings on the group, purely defined by their action. They have much in common with differential forms (which are in reality special cochains) but it should be kept in mind that here they are not necessarily acting on elements of a linear space. Here they take one, two, or more group elements to give numbers. The group elements have the role vectors have in the case of differential forms. Cochains may be defined on any group, even discrete ones—which is just the case of our interest. Here, α_0 is a zero-cochain, a function on parameter space whose value at point x is the phase $\alpha_0(x)$; α_1 is a one-cochain because it operates on one element g of G at point x of the parameter space to give $\alpha_1(x;g)$; a cochain taking two group elements as arguments will be a two-cochain, etc. An operation analogous to the exterior differentiation of differential forms is defined[8] on cochains: it is the derivative operation δ taking a p-cochain α_p into a $(p+1)$-cochain β_{p+1} according to[13]

$$\delta:\alpha_p\rightarrow \text{some } \beta_{p+1}=\delta\alpha_p,$$

$$\begin{aligned}\delta\alpha_p(x;g_1,g_2,\ldots,g_{p+1})&=\alpha_p(xg_1;g_2,\ldots,g_{p+1})-\alpha_p(x;g_1g_2,\ldots,g_{p+1})\\&\quad+\cdots+(-)^{p+1}\alpha_p(x;g_1,g_2,\ldots,g_p).\end{aligned} \tag{4.6}$$

An important property is the Poincaré lemma $\delta^2=0$, which can be verified directly from this expression. The first examples are

$$\delta\alpha_0(x;g)=\alpha_0(xg)-\alpha_0(x); \tag{4.7}$$

$$\delta\alpha_1(x;g,h)=\alpha_1(xg;h)-\alpha_1(x;gh)+\alpha_1(x;g); \tag{4.8}$$

$$\delta\alpha_2(x;g,h,f)=\alpha_2(xg;h,f)-\alpha_2(x;gh,f)+\alpha_2(x;g,hf)-\alpha_2(x;g,h). \tag{4.9}$$

A cochain α_p satisfying $\delta\alpha_p=0$ is a *closed* p-cochain, or a p-cocycle, and a cochain α_p for which a cochain α_{p-1} exists such that $\alpha_p=\delta\alpha_{p-1}$ is *exact*, or a *coboundary* (or trivial cocycle). An exact cochain is automatically closed. We see that condition (4.4) means that α_1 is closed,

$$\delta\alpha_1(x;g,h)=0, \tag{4.10}$$

still another form of the homomorphic condition. As to (4.5), it says simply that α_1 is exact:

$$\alpha_1(x;g)=\delta\alpha_0(x;g). \tag{4.11}$$

Summing up, the composition rule (4.2) implies the closedness of α_1; if in addition α_1 is a derivative, a redefinition of the functions exists such that it simply disappears. When α_1 is closed but not exact, it cannot be eliminated but the representation is still equivalent to a linear representation. A pure projective representation appears when, instead of (4.2), we only require

$$U(h)U(g)|\psi_x\rangle=\mathbf{e}^{i\alpha_2(x;g,h)}U(gh)|\psi_x\rangle, \tag{4.12}$$

allowing the composition to depend on the "position" x through a phase factor. The mapping $U:g\rightarrow U(g)$ is no more a homomorphism. Applying (4.3) successively, we have

$$\begin{aligned}&U(gh)|\psi_x\rangle=\mathbf{e}^{i\alpha_1(x;gh)}|\psi_{xgh}\rangle,\\&U(h)U(g)|\psi_x\rangle\\&\quad=\mathbf{e}^{i\alpha_1(x;g)}\mathbf{e}^{i\alpha_1(xg;h)}|\psi_{xgh}\rangle\\&\quad=e^{i[\alpha_1(xg;h)-\alpha_1(x;gh)+\alpha_1(x;g)]}U(gh)|\psi_x\rangle.\end{aligned} \tag{4.13}$$

Consequently,

$$\delta\alpha_1(x;g,h)=\alpha_2(x;g,h). \tag{4.14}$$

In this case α_1 is not closed and the representation is no more equivalent to a linear one. The cochain α_2 is an *obstruction* to homomorphism. On the other hand, ray representations like (4.3) require α_2 to be exact.

Let us see what comes out from the imposition of associativity: equaling

$$\begin{aligned}&U(f)[U(h)U(g)]|\psi_x\rangle\\&\quad=\mathbf{e}^{i\alpha_2(x;g,h)}U(f)U(gh)|\psi_x\rangle\\&\quad=\mathbf{e}^{i\alpha_2(x;g,h)}\mathbf{e}^{i\alpha_2(x;gh,f)}U(ghf)|\psi_x\rangle\end{aligned} \tag{4.15}$$

and

$$\begin{aligned}&[U(f)U(h)][U(g)|\psi_x\rangle]\\&\quad=\mathbf{e}^{i\alpha_2(xg;h,f)}U(hf)U(g)|\psi_x\rangle\\&\quad=\mathbf{e}^{i\alpha_2(xg;h,f)}\mathbf{e}^{i\alpha_2(x;g,hf)}U(ghf)|\psi_x\rangle\end{aligned} \tag{4.16}$$

brings forth, from (4.9), just the closedness of α_2,

$$\delta\alpha_2=0. \tag{4.17}$$

This "associativity condition" is of course coherent with (4.14).

Condition (4.14) has an interesting consequence. Suppose it holds and let us proceed to a redefinition of the operators U: define new operators U^* by

$$U^*(g)=\mathbf{e}^{-i\alpha_1(x;g)}U(g). \tag{4.18}$$

They depend, through the phase, on the point x at which they will operate and are, in this sense, "gaugefied" versions of the previous $U(g)$. In terms of such operators, (4.13) becomes

$$U^*(h)U^*(g)|\psi_x\rangle=U^*(gh)|\psi_x\rangle, \tag{4.19}$$

which is just of the form (4.2).

Concerning only the group operator representatives (and not the particular carrier space), it is expression (4.12) which characterizes a projective representation. Associativity implies that α_2 is a cocycle. If it is also exact, there exists a α_1 satisfying (4.14) which will appear as a ket phase and α_2 can be absorbed by the procedure just described into the "gaugefied" operators, in terms of which the representation reduces (but only locally in parameter space) to a linear one. We will say in this case that the representation is locally linear but globally projective. The unitary quantum operators to be studied in next section will be of this type. If α_2 is closed but not exact, there exists no α_1 as in (4.14) and α_2 cannot be eliminated. The projective representation is not even locally equivalent to a linear representation and is not of the form (4.3). Consequently, it is better to reserve the name "ray representations" to locally linear representations.

If an exact cochain $\delta\beta_1$ is added to α_2, the exact part can be eliminated but the nonexact "core" cannot. Adding an exact cochain is an equivalence relation, the corresponding classes being the elements of the quotient space of the closed by the exact cochains. This quotient space is the additive cohomology group $H^2(G)$. There is a one-to-one relation between the inequivalent projective representations and the elements of $H^2(G)$, which thereby "classifies" them.[7,8]

To obtain condition (4.17), we have taken associativity for granted in its usual way. If we are enough of a free thinker to accept that it holds up only to a phase factor,

$$[U(f)U(h)]U(g)|\psi_x\rangle = e^{i\alpha_3(x;g,h,f)}U(f)[U(h)U(g)]|\psi_x\rangle \tag{4.20}$$

then

$$\delta\alpha_2=\alpha_3 \tag{4.21}$$

instead of (4.17). Here, α_3 is a three-cochain, as it takes three elements of G to give the number $\alpha_3(x;g,h,f)$. When it is nonvanishing, α_2 is no more a cocycle and there is no associativity: α_3 is an obstruction to associativity. In principle, we can proceed with such successive steps of requirements and a corresponding hierarchy of closed and exact cochains. Nevertheless, associativity is part of the definition of a group and so desirable a property for a representation that it is usual to stop at this point. We say then simply that α_3 is an obstruction to the construction of projective representations.

It is also possible to introduce a notion akin to the interior product: Given the p-cochain α_p, its "interior product" with $h\in G$ is that $(p-1)$-cochain $\iota_h\alpha_p$ satisfying

$$[\iota_h\alpha_p](x;g_1,g_2,\ldots,g_{p-1})=\alpha_p(x;h,g_1,g_2,\ldots,g_{p-1}), \tag{4.22}$$

for all $g_1,g_2,\ldots,g_{p-1}$. A natural further step is to introduce a formal "Lie derivative" with respect to a $h\in G$ by

$$\lambda_h=\delta\circ\iota_h+\iota_h\circ\delta. \tag{4.23}$$

Some of its formal properties, again analogous to those of differential forms, are

$$\delta\circ\lambda_h=\delta\circ\iota_h\circ\delta=\lambda_h\circ\delta; \tag{4.24a}$$

$$(\lambda_h\alpha_0)(x)=\alpha_0(xh)-\alpha_0(x)=\delta\alpha_0(x;h); \tag{4.24b}$$

$$(\lambda_h\alpha_1)(x;g)=\alpha_1(xg;h)-\alpha_1(x;hg)+\alpha_1(xh;g); \tag{4.24c}$$

$$(\lambda_h\alpha_2)(x;g,j)=\alpha_2(xh;g,j)-\alpha_2(x;hg,j)+\alpha_2(xg;h,j). \tag{4.24d}$$

The limited character of such analogies should however be stressed. Unlike differential forms, the above cochains are not acting on a linear space and consequently share with them only some of their properties. They lack a tensorial character and, as a consequence, all the qualities coming with it. For example, there are no basis in terms of which any p-cochain can be written.

V. THE FUNDAMENTAL COCYCLE

As said in Sec. III, it is the toroidal lattice formed by the labels (m,n) of Schwinger's operators S_{mn} that constitute quantum phase space. Our objective, to which we finally arrive, is to show that indeed a certain cocycle (α_2 below) exists which endows the space of a structure similar to the symplectic structure of classical phase space and tends to the symplectic form in the classical limit. Consider the unitary operators of Sec. III. It comes directly from (3.12) and (4.3) that

$$\alpha_1(k;S_{mn})=(\pi/N)[(2k+m)n], \tag{5.1}$$

of which two particular cases are

$$\alpha_1(k;V)=(2\pi/N)k \tag{5.2}$$

and

$$\alpha_1(k;U)=0. \tag{5.3}$$

We need the two expressions

$$\alpha_1(kS_{mn};S_{rs})=\alpha_1(k+m;S_{rs})=(\pi/N)[2(k+m)+r]s, \tag{5.4}$$

and

$$\alpha_1(k;S_{mn}S_{rs})=(\pi/N)[(2k+m+r)(n+s)], \tag{5.5}$$

to verify, using (4.8), that

$$\delta\alpha_1(k;S_{mn},S_{rs})=\alpha_1(kS_{mn};S_{rs})-\alpha_1(k;S_{mn}S_{rs})+\alpha_1(k;S_{mn})=(\pi/N)[ms-nr]. \tag{5.6}$$

This is nonvanishing in general, hinting, after the discussion of the previous section, to a globally projective character. Indeed, from (3.13) we obtain

$$\alpha_2(k;S_{mn},S_{rs})=(\pi/N)[ms-nr], \tag{5.7}$$

so that α_2 is exact:

$$\alpha_2(k;S_{mn},S_{rs})=\delta\alpha_1(k;S_{mn},S_{rs}), \tag{5.8}$$

for any pair S_{mn},S_{rs}. This means that the representation only reduces to a linear one if we want to pay the price of "gaugefying" it as in (4.18): it is a ray representation, locally linear although globally projective.

Notice that $\alpha_2(k;S_{mn},S_{rs})$ is independent of the state label k. A particular value of interest is

$$\alpha_2(k;U,V)=\alpha_2(k;S_{10},S_{01})=\pi/N. \tag{5.9}$$

That the cochain α_2 is a cocycle is a consequence of the associativity condition (3.17):

$$\delta\alpha_2(k;S_{mn},S_{rs},S_{kl})=\alpha_3(k;S_{mn},S_{rs},S_{kl})=0. \tag{5.10}$$

Of course, this was already implied by the triviality (5.8) of α_2 and actually contained in the product rules (3.13). We should call attention to an obvious but important aspect. The cochains act on group elements to produce phases, exponentiated numbers. Unitary operators are not observables, only their Hermitian exponents are. The parameters m, n appear always exponentiated also, as in (3.9) and in the continuum limit [as in (3.19)] they do seem to tend to observables with classical counterparts. It is as if the parameters belonged not to the group but to its algebra. We must consequently be prepared to the fact that the relation between α_2 and Ω is exponential and, for facility, compare the results of their respective actions. There is no obvious correlation between associativity and the property related to the closedness of Ω, the Jacobi identity (2.18) for the Poisson bracket. Associativity is a much more general condition, a property of every group while Jacobi identity, typically an integrability condition, appears (exponentiated, as a property of the generators) only for Lie groups. Presumably this general property gets somehow weakened in the limiting process. An analogy may, however, help to shed some light on this point. There is a strong similarity of the formalism above with the basic structure of gauge theories: α_1 recalls the gauge potential A, δ the covariant derivative D, α_2 the field strength $F=DA$. Or, it happens that in gauge theories the closedness of F, $DF=0$ (the Bianchi identity) is precisely equivalent to the Jacobi identity for the gauge group generators.[14] We might conjecture that the closedness of α_2 is somehow related to that of Ω.

It is instructive to consider on the parameter space of the numbers m, n, r, etc. column vectors

$$X_{mn}=\sqrt{\frac{\pi}{N}}\begin{pmatrix}m\\ n\end{pmatrix},\quad X_{rs}=\sqrt{\frac{\pi}{N}}\begin{pmatrix}r\\ s\end{pmatrix},\quad \text{etc.},$$

with them as components. The row vectors X^T_{mn}, X^T_{rs} will behave as dual vectors by simple scalar product. Then, with the usual product of rows, matrices, and columns,

$$\alpha_2(k;S_{mn},S_{rs})=(\pi/N)\,[ms-nr]=X^T_{mn}\Omega X_{rs}, \tag{5.11}$$

where Ω is the symplectic matrix (2.16). On the toroidal grid formed by the parameters $\alpha_2(k;S_{mn},S_{rs})$ is proportional to the "area" defined by the vectors (m,n) and (r,s), as was the case for Ω in (2.17). We may also check that $\iota_{S_{mn}}\alpha_2$ is closed and takes a column vector X_{mn} into $X_{-n,m}$,

$$[\iota_{S_{mn}}\alpha_2](k;S_{rs})=\frac{\pi}{N}(-n,m)\begin{pmatrix}r\\ s\end{pmatrix}. \tag{5.12}$$

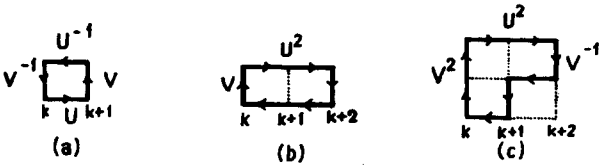

FIG. 1. The simplest loops on QPS: (a) an elementary loop; (b) a double loop with negative sense; (c) a triple loop. Each enclosed elementary cell contributes a phase $2\pi/N$ to S_{mn}.

This duality corresponds to relation established by Ω between vectors and forms. Furthermore, putting together the considerations on the continuum limit at the end of Sec. III and Eqs. (2.14) and (2.15), we see that α_2 plays on the lattice torus a role quite analogous to that of the symplectic form: from (3.21), we see that in the continuum limit α_2 gives (minus) half the value (2.15) of Ω applied to the corresponding vectors:

$$\alpha_2(k;S_{ab},S_{cd})=-\tfrac{1}{2}[ad-cb]=-\tfrac{1}{2}\Omega(J_{ab},J_{cd}).$$

Using (4.23) we find that

$$(\lambda_{S_{mn}}\alpha_2)(k;S_{rs},S_{pq})=0, \tag{5.13}$$

for all S_{mn}, S_{rs}, and S_{pq}, stating the invariance of α_2 under all transformations of the Weyl group. In this sense, all of them are "canonical transformations." Another analogy, trivial to obtain but interesting, comes from the very definition of α_2: It vanishes when applied to two commuting elements, just as Ω vanishes when applied to two fields corresponding to dynamical functions whose Poisson bracket vanishes. Such two fields are tangent to the same Lagrange manifold. On QPS, this corresponds to subsets of intercommuting operators. Finally, from (5.8), we see that the role of the canonical form σ is played by the cochain α_1.

Points in QPS can be attained from each other by successive applications of the operators U and V. Operators S_{mn} will meanwhile acquire phases. This is better seen if we start with some state $|v_k\rangle$ and look such successive transformations as forming paths on QPS. Each time U is applied the state is shifted and each time V is applied the ket gains a phase. This phase depends on the state arrived at. In Fig. 1(a), operator V acts at "$k+1$", but its inverse v^{-1} acts at "k." As a consequence of this point dependence, closed loops give a net result phase. Going around the loop in Fig. 1(a), for example, will give to $|v_k\rangle$ a phase $\epsilon^2=(2\pi/N)$. This ϵ^2 is the unit phase: It comes each time a unit cell in QPS is surrounded. The sum of phases is algebraic: Going around the unit loop in the inverse sense changes its sign. In our convention, positive sign is given by counterclockwise motion. So, the path of Fig. 1(b) contributes a phase $(-2\epsilon^2)$, that of Fig. 1(c) a phase $(-3\epsilon^2)$, etc. Closed loops may give vanishing phases. This is trivial for the two closed paths generated by U^N and V^N, which simply close around the torus, but there are nontrivial cases: In Fig. 2, the contributions from the two unit loops cancel each other. As α_2 measures just (half) the areas in units of ϵ, there is at work here a version of

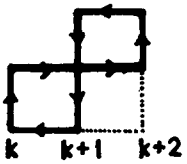

FIG. 2. A loop giving a vanishing contribution to the phase of S_{mn}.

Gauss theorem: the total (algebraic) area circumvented by a loop is obtained by just following the loop step by step, at each step summing the corresponding α_1, as given by (5.2) and (5.3). As exhibited in Fig. 3, the product $S_{mn}S_{rs}$ is equivalent to taking S_{mn} then S_{rs} only when $\alpha_2(k;S_{mn},S_{rs}) = 0$. The closed paths of Figs. 1 and 2 are projections of paths in the space of the operators S_{mn}, where the paths are, as a rule, open. A kind of nonintegrability appears: Starting from a given point, the phase at another point will depend on the path, unless the "flux" of $\mathbf{a}_2$ through the surface defined by any two paths is zero. In this sense α_1 is a nonintegrable phase like those of gauge theories[15] and α_2 would act as the corresponding "curvature." As already mentioned, there are many aspects in common with gauge theories in the present formalism, but we shall not discuss them here. Neither shall we consider the possible relation of α_1 to a generalized[16] Berry's phase,[17] a subject deserving further study.

As an example, the commutator $V^{-1}U^{-1}VU$ of Fig. 1(a) produces in operator space (see Fig. 4) an arc which fails to close precisely by the phase ϵ^2. Such trajectories in operator space only close when the unit cell is surrounded a multiple of N times, in which case it becomes a closed spiral. The role of α_2, similar to a curvature on QPS, is different here: As it measures such defects in the operator space, it is reminiscent of that of torsion in differential geometry.

In the continuum limit we must consider "large" regions of sizes $m\epsilon$ and $n\epsilon$ tending to limits a and b and the operators (now putting $\hbar$ back into our expressions $U^m \to e^{iap/\hbar}$ and $V^n \to e^{ibq/\hbar}$. The phase

$$\alpha_2(k;U^m,V^n) = (\pi/N)mn = (\epsilon^2/2)mn$$

tends to $\frac{1}{2}ab$, just (half) the value of $\Omega(aX_q,bX_p)$. Actually, to examine the continuum limit, as well as to get some more insight on the role of the cocycle α_2 it is convenient to apply the formula giving the Weyl–Wigner transform $W(\mathbf{AB}) = (\mathbf{AB})_W$ of the product of two operators $\mathbf{A}$ and $\mathbf{B}$ in terms of their transforms $W(\mathbf{A}) = \mathbf{A}_W(q,p)$ and $W(\mathbf{B}) = \mathbf{B}_W(q,p)$, which is

$$W(\mathbf{AB}) = e^{i(\hbar/2)[\partial^A{}_q\partial^B{}_p - \partial^A{}_p\partial^B{}_q]}\mathbf{A}_W(q,p)\mathbf{B}_W(q,p). \tag{5.14}$$

The upper indices in $\partial^A{}_q$, $\partial^B{}_p$ are reminders: $\partial^A{}_p$ is the derivative with respect to q but which applies only on $\mathbf{A}_W(q,p)$; $\partial^B{}_p$ derives with respect to p but only acts on $\mathbf{B}_W(q,p)$, etc. The Poisson bracket always comes up at first order in $\hbar$:

$$W(\mathbf{AB}) = \mathbf{A}_W(q,p)\mathbf{B}_W(q,p) - (\hbar/2i) \times\{\mathbf{A}_W(q,p),\mathbf{B}_W(q,p)\} + \cdots \tag{5.15}$$

but the Weyl–Wigner transformed functions $\mathbf{A}_W(q,p)$ and $\mathbf{B}_W(q,p)$ may still exhibit additional powers of $\hbar$, depending on their explicit form in terms of q and p. In fact, only in the strict classical ($\hbar \to 0$) limit will such functions reduce to their classical counterparts. Getting the Poisson bracket from a quantum commutator is only achieved when we pass from a noncommutative algebra to a commutative one at the price of ignoring the cell structure of quantum phase space.[18] Only then $(i/\hbar) \times[\mathbf{A},\mathbf{B}] \to \{A_{\text{clas.}},B_{\text{clas.}}\}$. To clarify this point, let us consider the operators $\mathbf{A} = S_{a0} = e^{(i/\hbar)aq}$ and $\mathbf{B} = S_{0b} = e^{(i/\hbar)bp}$. From the previous formalism, their product will be

$$\mathbf{AB} = e^{i\alpha_2(\mathbf{A},\mathbf{B})}S_{ab} = e^{i\alpha_2(\mathbf{A},\mathbf{B})}e^{(i/\hbar)(aq+bp)}. \tag{5.16}$$

The Weyl–Wigner transform of the right-hand side is

$$W(\mathbf{AB}) = e^{i\alpha_2(\mathbf{A},\mathbf{B})}e^{(i/\hbar)(aq+bp)}, \tag{5.17}$$

where now q and p behave like classical variables. On the other hand, (5.14) will say that

$$W(\mathbf{AB}) = e^{i(\hbar/2)[\partial^A{}_q\partial^B{}_p - \partial^A{}_p\partial^B{}_q]}[e^{(i/\hbar)(aq+bp)}]. \tag{5.18}$$

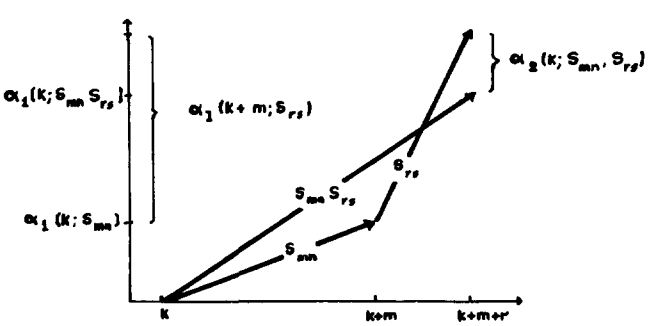

FIG. 3. The fundamental cocycle measures the phase difference between $S_{mn}S_{rs}$ and the successive applications of S_{mn} and S_{rs}.

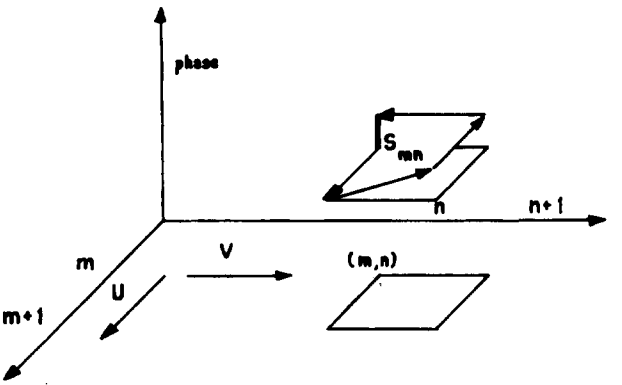

FIG. 4. An elementary loop in parameter space corresponds to an open trajectory in operator space.

We see that in some way α_2 sums up all the intricate action of the exponentiated operator. The present example is specially simple but reflects much of the fundamental structure of the continuum quantum phase space, as in this case S_{ab} is a typical base element. The Poisson bracket is constant and it is possible to write down the exact result,

$$W(\mathbf{AB}) = e^{-(i/2\hbar)ab} e^{(i/\hbar)(aq+bp)}, \tag{5.19}$$

so that

$$\alpha_2(\mathbf{A},\mathbf{B}) = -(1/2\hbar)\{\mathbf{A},\mathbf{B}\}. \tag{5.20}$$

An analogous result would come if we took operators of type (3.21). In such cases related to the harmonic oscillator, whose semiclassical approximation is exact, α_2 gives the classical result up to a factor $\hbar^{-1}$. This is indeed the hallmark of the quantum structure of phase space embodied in α_2, which is not at all a classical object. It is expressed above in terms of the Poisson bracket, but of Weyl–Wigner representatives of quantum objects. In this continuum case, α_2 heralds the noncommutativity of the basic pair **q–p**. In the general case, it highlights the fundamental cellular structure of QPS.

VI. SUMMARY

Every feature of classical mechanics stems from some quantum mechanical feature. Let us try to review the analogies and differences between the cocycle α_2 and the symplectic form. To begin with, Ω is globally defined on the classical state space and $\alpha_2(k;S_{mn},S_{rs})$ is independent of the state label k. The first is invariant under canonical transformations, the second under all unitary transformations. Both measure areas defined by vectors in the corresponding spaces. The closedness of Ω guarantees the Jacobi identity for the Poisson brackets, that of α_2 the projective character of the Weyl representation. Classical Lagrange manifolds are on QPS replaced by subsets of intercommuting unitary operators. The symplectic form is a linear operator, which we could not expect of α_2. Finally, α_2 tends to the symplectic form Ω when, in the continuum limit, the noncommutativity of dynamical variables is relaxed. The cocycle α_2 is that feature of quantum mechanics on which the symplectic structure of classical mechanics casts its roots.

ACKNOWLEDGMENTS

This work was partially supported by FINEP, Rio de Janeiro. The authors would like to thank the CNPq, Brazil, who gave them partial support.

[1] D. Galetti and A. F. R. de Toledo Piza, Physica A **149**, 267 (1988); W. K. Wooters, Ann. Phys. **176**, 1 (1987); O. Cohendet, Ph. Combe, M. Sirugue, and M. Sirugue-Collin, J. Phys. A **21**, 2875 (1988).
[2] J. Zak, J. Math. Phys. **30**, 1591 (1989); J. Zak, Phys. Rev. B **39**, 694 (1989).
[3] J. Schwinger, Proc. Natl. Acad. Sci. USA **46**, 570, 893 (1960); **47**, 1075 (1961). See also J. Schwinger, *Quantum Kinematics and Dynamics* (Benjamin, New York, 1970).
[4] V. Arnold, *Mathematical Methods in Classical Mechanics* (Springer, New York, 1978).
[5] R. Abraham and J. E. Marsden, *Foundations of Mechanics* (Benjamin–Cummings, Reading, MA, 1978), 2nd. ed.
[6] M. Hamermesh, *Group Theory and its Applications to Physical Problems* (Addison-Wesley, Reading, MA, 1962).
[7] L. Michel, in *Group Theoretical Concepts and Methods in Elementary Particle Physics*, edited by F. Gursey, Istambul Summer School of Theoretical Physics 1963 (Gordon and Breach, New York, 1964).
[8] A. A. Kirillov, *Elements of the Theory of Representations* (Springer, Berlin, 1976).
[9] J. E. Moyal, Proc. Cambridge Phys. Soc. **41**, 99 (1949).
[10] H. Weyl, *The Theory of Groups and Quantum Mechanics* (Dover, New York, 1950), Chap. 4, Sec. 14.
[11] C. Isham, *Relativity, Groups and Topology II*, edited by B. de Witt and R. Stora, Les Houches Summer School 1983 (North-Holland, Amsterdam, 1984).
[12] G.'t Hooft, Nucl. Phys. B **138**, 1 (1978).
[13] L. D. Faddeev and S. L. Shatashvilli, Theor. Math. Phys. **60**, 770 (1984).
[14] Y. M. Cho, J. Math. Phys. **16**, 2029 (1975).
[15] T. T. Wu and C. N. Yang, Phys. Rev. D **12**, 3845 (1975).
[16] Y. Aharonov and J. Anandan, Phys. Rev. Lett. **58**, 1593 (1987).
[17] M. V. Berry, Proc. R. Soc. London Ser. A **392**, 45 (1984).
[18] S. R. de Groot and L. G. Suttorp, *Foundations of Electrodynamics* (North-Holland, Amsterdam, 1972), Chap. VI and the appendix.

On boson condensation considering a generalized Casimir example

K. Voigt
Sektion Mathematik der Universität Leipzig, Augustusplatz 10, Leipzig 7010, German Democratic Republic

(Received 12 October 1989; accepted for publication 11 July 1990)

From the concept of generalized condensation [M. Van den Berg, J. T. Lewis, and J. V. Pulè, Helv. Phys. Acta **59**, 1271 (1986)] it is known that two critical densities ρ_c and ρ_m exist for a free boson gas. Density ρ_c is the classical one and ρ_m is the critical density below which there can be no macroscopic occupation of ground state. A free boson gas is studied in a weak external potential which behaves asymptotically like $|x_1|^{\alpha_1} + |x_2|^{\alpha_2} + \cdots + |x_d|^{\alpha_d}$ near the origin. It is shown that there are only two possibilities to get $\rho_c < \rho_m < \infty$, namely, $\alpha_1 = \alpha_2 = \infty$ and $d \geqslant 3$ (this corresponds to Dirichlet boundary conditions), and $\alpha_1 = 2$ and $d \geqslant 2$ (i.e., a harmonic oscillator).

I. INTRODUCTION

The aim of this paper is twofold: to give some more examples of generalized condensation[1] of the free boson gas and to simplify the proofs of already known examples[2-5] of macroscopic occupation of low lying levels. Van den Berg, Lewis, and Pulè gave a unified treatment of Bose–Einstein condensation in noninteracting systems of bosons with a general single-particle Hamiltonian. They show that there are two critical densities: ρ_c, at which singularities in the thermodynamic functions occur, and ρ_m, below which there can be no macroscopic occupation of the ground state. To the best of our knowledge there is only one example given by van den Berg[3] which shows that $\rho_c < \rho_m < \infty$ can occur. Van den Berg considers a prism of sides $L_1, L_2, \ldots, L_d, d \geqslant 3$, and takes the single-particle Hamiltonian to be the Laplacian with Dirichlet boundary conditions. This is one generalization of Casimir's[6] examples, and will be a special case of our setting. Consider as a single-particle Hamiltonian,

$$H_{L_1,L_2,\ldots,L_d} = -\Delta + V_1\left(\frac{x_1}{L_1}\right) + \cdots + V_d\left(\frac{x_d}{L_d}\right), \quad (1)$$

such that trace $(e^{-\beta H_{L_1,\ldots,L_d}}) < \infty$ for all β in $(0,\infty)$. Let

$$E_k(L_1,L_2,\ldots,L_d) = E_{k_1}(L_1) + E_{k_2}(L_2) + \cdots + E_{k_d}(L_d), \quad k = (k_1,\ldots,k_d)$$

be the eigenvalues of $H_{L_1,L_2,\ldots,L_d}$, where $E_1(L_i) < E_2(L_i) \leqslant \cdots$ are the eigenvalues of $-d^2/dx^2 + V_i(x/L_i)$ in ascending order.

Note that the thermodynamic functions as a function of the total mean density do not depend directly on the eigenvalues but only on the distances

$$\lambda_k(L_1,\ldots,L_d) = E_k(L_1,\ldots,L_d) - E_{(1,1,\ldots,1)}(L_1,\ldots,L_d), \quad k = (k_1,\ldots,k_d).$$

It is convenient to introduce the distribution functions

$$F_{L_1,\ldots,L_d}(\lambda) = \frac{1}{L_1 \cdot \ldots \cdot L_d} \operatorname{Card}\{k : \lambda_k(L_1,\ldots,L_d) \leqslant \lambda\}.$$

To discuss condensation it is necessary to introduce first the occupation numbers $n_k(L_1,\ldots,L_d)$ of the k th energy level of $H_{L_1,\ldots,L_d}$. Then one takes the thermodynamic limit $L_1, L_2, \ldots, L_d \to \infty$ while one keeps the density ρ fixed. In the following sections we shall be interested in how this limit must be taken. However for this starting section this will be not important we simply write $L \to \infty$:

$$n_k(L_1,\ldots,L_d) = \frac{1}{L_1 \cdot \cdots \cdot L_d} \cdot \frac{z(L_1,\ldots,L_d)}{e^{\beta\lambda_k(L_1,\ldots,L_d)} - z(L_1,\ldots,L_d)},$$

where the activity $z(L_1,\ldots,L_d)$ is the unique solution of

$$\rho = \sum_k n_k(L_1,\ldots,L_d) = \int_{[0,\infty)} \frac{z(L_1,\ldots,L_d)}{e^{\beta\lambda} - z(L_1,\ldots,L_d)}\, dF_{L_1,\ldots,L_d}(\lambda), \quad (2)$$

with fixed density ρ. Define

$$\Phi_{L_1,\ldots,L_d}(\beta) = \int_{[0,\infty)} e^{-\beta\lambda}\, dF_{L_1,\ldots,L_d}(\lambda).$$

Van den Berg, Lewis, and Pulè[1] derived the following result: If $\Phi(\beta) = \lim_{L\to\infty} \Phi_{L_1,\ldots,L_d}(\beta)$ exists for all β in $(0,\infty)$ and is nonzero for some β in $(0,\infty)$ then there exists a unique distribution function $F(\lambda)$ determined by

$$\Phi(\beta) = \int_{[0,\infty)} e^{-\beta\lambda}\, dF(\lambda).$$

One has for $\varepsilon > 0$

$$\lim_{L\to\infty} \int_{[\varepsilon,\infty)} \frac{z}{e^{\beta\lambda} - z}\, dF_{L_1,\ldots,L_d}(\lambda) = \int_{[\varepsilon,\infty)} \frac{z}{e^{\beta\lambda} - z}\, dF(\lambda), \quad \text{for all } z \in (0,1]. \quad (3)$$

Define now the critical density ρ_c by

$$\rho_c = \int_{[0,\infty)} \frac{1}{e^{\beta\lambda} - 1}\, dF(\lambda), \quad \text{if } (e^{\beta\lambda} - 1)^{-1} \text{ is integrable with respect to } F$$

$$\rho_c = \infty, \quad \text{otherwise.}$$

Then $z = \lim_{L\to\infty} z(L_1,\ldots,L_d)$ exists and equals 1 if $\rho \geqslant \rho_c$ and is the unique root in $(0,1)$ of the equation

$$\rho = \int_{[0,\infty)} \frac{z}{e^{\beta\lambda} - z}\, dF(\lambda) \quad \text{if } \rho < \rho_c.$$

 0022-2488/90/122996-07$03.00

If $F(\lambda) \sim C\lambda^{\alpha}$ as $\lambda \downarrow 0$ with $\alpha > 1$ then ρ_c is finite. In order to decide whether the k th energy level will be macroscopically occupied, i.e.,

$$n_k = \lim_{L_1,\ldots,L_d \to \infty} n_k(L_1,\ldots,L_d) > 0,$$

one needs some finer properties of the spectrum of $H_{L_1,\ldots,L_d}$, or in other words of $F_{L_1,\ldots,L_d}(\lambda)$ as $\lambda \downarrow 0$. To make now our setting precise we consider the single-particle Hamiltonian $H_{L_1,\ldots,L_d}$ of (1) acting on the Hilbert space $L^2(\mathbb{R}^d)$ and suppose[7] the existence of $\phi(\beta)$ described above such that (3) holds and $z = \lim_{L\to\infty} z(L_1,\ldots,L_d)$ exists. Suppose moreover that the potentials $V_i(x_i)$ of (1) are continuous functions defined on $\mathbb{R}$, and behave around their minima asymptotically like $a_i|x_i|^{\alpha_i}, i = 1,\ldots,d$, in the sense of the following definition.

Definition 1: Let $a,\alpha > 0$. If for every ε, $0 < \varepsilon < a$, there exists $x_\varepsilon > 0$ such that

$$(a-\varepsilon)\cdot|x|^{\alpha} \leqslant V(x) \leqslant (a+\varepsilon)|x|^{\alpha}, \quad \text{for all } x,\ |x| \leqslant x_\varepsilon \tag{4a}$$

$$(a-\varepsilon)x_\varepsilon^{\alpha} \leqslant V(x), \quad \text{otherwise.} \tag{4b}$$

Then the continuous function $V(x)$ is called a asymptotically $a\cdot|x|^{\alpha}$ potential.

II. ESTIMATES ON THE DISTRIBUTION FUNCTIONS $F_{L_1,\ldots,L_d}$

In what follows we use the notations of Reed and Simon.[8]

Definition 2: If A,B are two self-adjoint non-negative operators, densely defined in Hilbert spaces $\mathscr{H}, \mathscr{H}_1$, respectively, $\mathscr{H}_1 \subseteq \mathscr{H}$. We write $0 \leqslant A \leqslant B$ if and only if (i) $Q(A) \supseteq Q(B)$ $Q(.)$ the domain of definition of the forms. (ii) $0 \leqslant (\psi,A\psi) \leqslant (\psi,B\psi), \psi \in Q(B)$. Further we make extensive use of Dirichlet Neumann bracketing. Denote by $-\Delta_{\mathrm{D}}^{\Omega}$ (resp. $-\Delta_{\mathrm{N}}^{\Omega}$) the Dirichlet (resp. Neumann) Laplacian with Dirichlet (resp. Neumann) boundary conditions at the boundary of the open domain Ω. $N_{\mathrm{D}}(\Omega,E)$ [resp. $N_{\mathrm{N}}(\Omega,E)$] denote the number of eigenvalues $\leqslant E$ of these operators.

Let us first discuss the one-dimensional case, $d = 1$. Fix $\varepsilon > 0$, and let $\lambda_\varepsilon = (a-\varepsilon)x_\varepsilon^{\alpha}, L = L_1, \alpha = \alpha_1$. We have

$$H_L \leqslant -\Delta_{\mathrm{D}}^{(-Lx_\varepsilon, Lx_\varepsilon)} + (a+\varepsilon)|x/L|^{\alpha}. \tag{5}$$

For the right-hand side of (5) Van den Berg and Lewis[5] have already derived an upper bound, say $M_k(\varepsilon)L^{-2\alpha/(\alpha+2)}$ for the eigenvalues. Hence it holds that

$$E_k(L) \leqslant M_k(\varepsilon)L^{-2\alpha/(\alpha+2)} \tag{6}$$

for the eigenvalues $E_1(L) < E_2(L) \leqslant \cdots$ of H_L. In Appendix A we will calculate also a lower bound for the distance of the two first eigenvalues:

$$E_2(L) - E_1(L) \geqslant mL^{-2\alpha/(\alpha+2)} \tag{7}$$

with some positive constant m.

Lemma 1: There exist sequences $\{A_n\}_{n\in\mathbb{N}}, \{B_n\}_{n\in\mathbb{N}}$ such that for all $\lambda < \lambda_\varepsilon$

$$-\frac{n}{L} + \frac{A_n}{\pi(a+\varepsilon)^{1/\alpha}}\cdot\lambda^{1/2+1/\alpha} \leqslant F_L(\lambda) \leqslant \frac{B_n}{\pi(a-\varepsilon)^{1/\alpha}}(\lambda + M_1L^{-2\alpha/(\alpha+2)})^{1/2+1/\alpha} + \frac{n+2}{L}, \tag{8a}$$

for sufficiently large L, and

$$\lim_{n\to\infty} A_n = \lim_{n\to\infty} B_n = \int_{[-1,1]} (1-|t|^{\alpha})^{1/2}\,dt = \frac{2\cdot\Gamma(\frac{3}{2})\cdot\Gamma(1/\alpha)}{\alpha\cdot\Gamma(\frac{3}{2}+1/\alpha)}, \tag{8b}$$

where $\Gamma(\cdot)$ denotes the usual gamma function.

Proof: From (6) and the definition of F_L one has

$$(1/L)\cdot\mathrm{Card}\{k{:}E_k(L) \leqslant \lambda\} \leqslant F_L(\lambda) \leqslant (1/L)\ \mathrm{Card}\{k{:}E_k(L) \leqslant E\},$$

where $E = \lambda + M_1L^{-2\alpha/(2+\alpha)}$.

If L is large enough to provide $E < \lambda_\varepsilon$ we make the following choices to prove the second relation of (8a) for $n = 2l$ (the case $n = 2l+1$ can be handled in a similar manner):

$$x_E = \left(\frac{E}{a-\varepsilon}\right)^{1/\alpha},$$

$$I_k = \left(\frac{k}{l}x_EL, \frac{k+1}{l}x_EL\right), \quad k = -l,\ldots,l-1,$$

$$I_l = (x_EL, x_\varepsilon L), \quad I_{l+1} = (x_\varepsilon L, \infty),$$

$$I_{-l-1} = (-x_\varepsilon L, -x_EL),$$

$$I_{-l-2} = (-\infty, -x_\varepsilon L),$$

$$\Omega = \bigcup_{k=-l-2}^{l+1} I_k.$$

By Dirichlet–Neumann bracketing[6] one gets

$$-\Delta + V\left(\frac{x}{L}\right) \geqslant \bigoplus_{k=-l-2}^{l+1}\left(-\Delta_N^{I_k} + V\left(\frac{x}{L}\right)\right). \tag{9}$$

If one substitutes now on each interval I_k the potential $V(x/L)$ by the constant potential equal to the minimum of $V(x/L)$ on I_k then one gets yet a smaller operator which has eventually more eigenvalues smaller than E. Obviously the minimum of $V(x/L)$ on I_l is E, whereas that on I_{l+1} equals λ_ε. As $\lambda_\varepsilon > E$ we have $N_{\mathrm{N}}(I_l,0) = N_{\mathrm{N}}(I_{-l-1},0) = 1$, and $N_{\mathrm{N}}(I_{l+1},E-\lambda_\varepsilon) = N_{\mathrm{N}}(I_{-l-2},E-\lambda_\varepsilon) = 0$ thus we get from (9)

$$F_L(\lambda) \leqslant \frac{2}{L}\sum_{k=0}^{l-1} N_{\mathrm{N}}\left(I_k, E-(a-\varepsilon)\left(\frac{k}{l}x_E\right)^{\alpha}\right) + \frac{2}{L} \leqslant \frac{2}{L}\sum_{k=0}^{l-1}\frac{x_EL}{\pi l}\cdot\left(E-\left(\frac{k}{l}\right)^{\alpha}E\right)^{1/2} + \frac{2l+2}{L} = \frac{E^{1/2+1/\alpha}}{\pi\cdot(a-\varepsilon)^{1/\alpha}}\cdot B_n + \frac{n+2}{L},$$

where

$$B_n = 2\sum_{k=0}^{l-1}\frac{1}{l}\left(1-\left(\frac{k}{l}\right)^{\alpha}\right)^{1/2}, \tag{10}$$

which is a lower Riemann sum of the integral in (8b). Let us sketch the proof of the first relation of (8a) for $n = 2l$. We

make the following choices:

$$E=\lambda,\quad x_E=\left(\frac{E}{x+\varepsilon}\right)^{\alpha},\quad \Omega=\bigcup_{k=-1}^{l-1} I_k,$$

I_k being unchanged. Instead of relation (9) we obtain

$$-\Delta+V\left(\frac{x}{L}\right)\leqslant\bigoplus_{k=-1}^{l-1}\left(-\Delta_D^{I_k}+V\left(\frac{x}{L}\right)\right). \tag{11}$$

Instead of the minimum we take the maximum of $V(x/L)$ on each of the intervals I_k. Here, A_n will be defined in a similar manner as B_n in (10) and turns out to be a upper Riemann sum of the integral in (8b). The lemma now follows.

Now let the dimension d be arbitrary. For $\varepsilon>0$ define $\lambda_\varepsilon=\min\{(a-\varepsilon)x_{i,\varepsilon}^{\alpha_i}: i=1,\ldots,d\}$, where $x_{i,\varepsilon}$ are defined by (4) corresponding to the potentials V_i.

Lemma 2: There exist numbers C_1, C_2 such that for all $\lambda<\lambda_\varepsilon$,

$$F_{L_1,\ldots,L_d}(\lambda)\leqslant C_1(\lambda+C_2(L_1^{-\gamma_1}+\cdots+L_d^{-\gamma_d}))^{d/2+1/\alpha_1+\cdots+1/\alpha_d}, \tag{12}$$

where

$$\gamma_i=2\alpha_i/(\alpha_i+2),\quad i=1,\ldots,d. \tag{13}$$

Proof: From Lemma 1 we have for $d=1, n=1$

$$F_{L_1}(\lambda)\leqslant\frac{2}{\pi(a_1-\varepsilon)^{1/\alpha_1}}(\lambda+M_1L_1^{-\gamma_1})^{1/2+1/\alpha}+\frac{L}{3}\leqslant C_1(\lambda+C_2L_1^{-\gamma_1})^{1/2+1/\alpha_1},$$

with

$$C_1=\frac{2\max(1,2^{1-\gamma_1})}{\pi(a_1-\varepsilon)^{1/\alpha_1}},\quad C_2=M_1+\left(\frac{3\pi(a_1-\varepsilon)^{1/\alpha_1}}{2}\right)^{\gamma_1}.$$

The second inequality is a consequence of $a^\gamma+b^\gamma\leqslant\max(1,2^{1-\gamma})(a+b)^\gamma$. The lemma follows now by simple induction.

III. CONDENSATION IN THE PRESENCE OF $V(x)$

As already remarked in the introduction ρ_c is finite whenever $F(\lambda)\sim C\lambda^{\alpha}$ as $\lambda\downarrow 0$ with $\alpha>1$.

Lemma 3: If $V(x)\geqslant 0$ is continuous, $V(x)\to\infty$ as $|x|\to\infty$, $x\in\mathbb{R}^d$ and if $E_1(L_1,\ldots,L_d)<E_2(L_1,\ldots,L_d)\leqslant\cdots$ are the ordered eigenvalues of the Hamiltonian

$$H_{L_1,\ldots,L_d}=-\Delta+V\left(\frac{x_1}{L_1},\frac{x_2}{L_2},\ldots,\frac{x_d}{L_d}\right) \tag{14}$$

and if $\lim_{L\to\infty}E_1(L_1,\ldots,L_d)=0$ there exists

$$F(\lambda)=\lim_{L\to\infty}F_{L_1,\ldots,L_d}(\lambda)=\frac{\tau_d}{(2\pi)^d}\int_{V(x)<\lambda}(\lambda-V(x))^{d/2}\,dx, \tag{15}$$

τ_d being the volume of the unit ball in $\mathbb{R}^d$.

This is a slight generalization[9] of Pulè's Lemma 1 in Ref. 10 already discussed by Voigt.[11] The proof is very similar to our Lemma 1 and will therefore be omitted. We remark that the operator (14) is quite more general than our $H_{L_1,\ldots,L_d}$ introduced at the end of the introduction. Let us return now to the latter.

Proposition 1: If $V(x)=V_1(x_1)+\cdots+V_d(x_d)$, each of the V_i being a asymptotically $a_i|x_i|^{\alpha_i}$ potential as described in Definition 1, then

$$C^-(a,\alpha,\varepsilon)\lambda^{d/2+1/\alpha_1+\cdots+1/\alpha_d}\leqslant F(\lambda)\leqslant C^+(a,\alpha,\varepsilon)\lambda^{d/2+1/\alpha_1+\cdots+1/\alpha_d} \tag{16}$$

for all $\lambda<\lambda_\varepsilon$, where

$$C^{\pm}(a,\alpha,\varepsilon)=\pi^{-d/2}\frac{(\alpha_1\cdot\cdots\cdot\alpha_d)^{-1}}{(a_1+\varepsilon)^{1/\alpha_1}\cdot\cdots\cdot(a_d\pm\varepsilon)^{1/\alpha_d}}\times\frac{\Gamma(1/\alpha_1)\cdot\cdots\cdot\Gamma(1/\alpha_d)}{\Gamma(d/2+1/\alpha_1+\cdots+1/\alpha_d)}. \tag{17}$$

Proof: Define $W^{\pm}(x)=(a_1\pm\varepsilon)|x|^{\alpha_1}+\cdots+(a_d\pm\varepsilon)|x|^{\alpha_d}$.

If $\lambda<\lambda_\varepsilon$ then one can conclude

$$\frac{\tau_d}{(2\pi)^d}\int_{W^+(x)<\lambda}(\lambda-W^+(x))^{d/2}\,dx\leqslant\frac{\tau_d}{(2\pi)^d}\int_{W^+(x)<\lambda}(\lambda-V(x))^{d/2}\,dx\leqslant F(\lambda)$$

and

$$F(\lambda)\leqslant\frac{\tau_d}{(2\pi)^d}\int_{V(x)<\lambda}(\lambda-W^-(x))^{d/2}\,dx\leqslant\frac{\tau_d}{(2\pi)^d}\int_{W^-(x)<\lambda}(\lambda-W^-(x))^{d/2}\,dx.$$

But

$$\int_{\sum_{i=1}^d a_i|x_i|^{\alpha_i}<\lambda}\left(\lambda-\sum_{i=1}^d a_i|x_i|^{\alpha_i}\right)^{d/2}dx=\lambda^{d/2+1/\alpha_1+\cdots+1/\alpha_d}\frac{2^d(\alpha_1\cdot\cdots\cdot\alpha_d)^{-1}}{a_1^{1/\alpha_1}\cdot\cdots\cdot a_d^{1/\alpha_d}}\cdot\int_{\substack{t_1+\cdots+t_d<1\\ t_i>0}}(1-t_1-\cdots-t_d)^{d/2}t_1^{1/\alpha_1-1}\cdot\cdots\cdot t_d^{1/\alpha_d-1}\,dt_1\cdot\cdots\cdot dt_d$$

$$=\lambda^{d/2+1/\alpha_1+\cdots+1/\alpha_d}\frac{2^d(\alpha_1\cdot\cdots\cdot\alpha_d)^{-1}}{a_1^{1/\alpha_1}\cdot\cdots\cdot a_d^{1/\alpha_d}}\cdot\frac{\Gamma(d/2+1)\Gamma(1/\alpha_1)\cdot\cdots\cdot\Gamma(1/\alpha_d)}{\Gamma(d/2+1/\alpha_1+\cdots+1/\alpha_d)}.$$

Corollary: Under the assumptions of the Introduction it holds that $\rho_c < \infty$ if and only if $d/2 + 1/\alpha_1 + \cdots + 1/\alpha_d > 1$.

Proof: Obvious from (3) and Proposition 1.

Using the properties of the distribution functions $F_L(\lambda)$ we obtain in the following theorem results concerning the macroscopic occupation of the low-lying levels. Alternatively one can use the results of Ref. 1 if one studies the Laplace transforms $\Phi_L(\beta)$ and the limit of the rescaled density distributions $\gamma_L(u)$ defined there. Our concept is much more direct and seems to be simpler.

IV. MACROSCOPIC OCCUPATION OF LOW-LYING LEVELS

Theorem: Let the infinite volume[12] limit ($L \to \infty$) be such that the mean density ρ is kept fixed and

$$1.\ L_1^{\gamma_1} \geqslant L_2^{\gamma_2} \geqslant \cdots \geqslant L_d^{\gamma_d} \to \infty, \text{ where } \gamma_i = 2\alpha_i(\alpha_i + 2); \tag{18}$$

$$2.\ \lim_{L_d \to \infty} \frac{\log L_2}{L_3 \cdot \cdots \cdot L_d} = A \qquad (d \geqslant 3);$$

$$3.\ \lim_{L_d \to \infty} \frac{\log L_1}{L_2 \cdot \cdots \cdot L_d} = B \qquad (d \geqslant 2);$$

$$4.\ \lim_{L_d \to \infty} \frac{L_1}{L_2 L_3 \cdot \cdots \cdot L_d} = C \qquad (d \geqslant 2);$$

$$5.\ \lim_{L_d \to \infty} \frac{L_1^{(\alpha_1 - 2)/(\alpha_1 + 2)}}{L_2 L_3 \cdots L_d} = D \qquad (d \geqslant 2).$$

Let us differentiate between the following cases.

(*i*) *Case* $1/\alpha_1 + 1/\alpha_2 > 0$

$\alpha_1 < 2$: then only the ground state can be macroscopically occupied, $n_{1,1,\ldots,1} = (\rho - \rho_c)_+$.

$\alpha_1 = 2$: there exists a critical density ρ_m: $\rho_m = \rho_c + B/2\beta\sqrt{a_1}$, and none of the single-particle states are macroscopically occupied if $\rho \leqslant \rho_m$. If $\rho > \rho_m$ again only the ground state will be macroscopically occupied, $n_{1,1,\ldots,1} = \rho - \rho_m$.

$\alpha_1 > 2$: In dependence on D one has three cases:

$D = 0$: one has the same situation as for $\alpha < 2$, $n_{1,1,\ldots,1} = (\rho - \rho_c)_+$.

$0 < D < \infty$: If $\rho < \rho_c$ then $n_{k,1,1,\ldots,1} > 0, k = 1,2,\ldots$.

$D = \infty$: then none of the single-particle states are macroscopically occupied.

(*ii*) *Case* $1/\alpha_1 + 1/\alpha_2 = 0$

This means that there are Dirichlet boundary conditions in the first and the second direction. Obviously $d \geqslant 3$ is necessary to have condensation. There exists a second critical density $\rho_m = \rho_c + (2 \cdot A)/(\pi \cdot \beta)$. If $\rho \leqslant \rho_m$ none of the single-particle states are macroscopically occupied. If $\rho > \rho_m$ one has to differentiate between three cases in dependence of C.

$C = 0$: only the ground state will be macroscopically occupied, $n_{1,1,\ldots,1} = \rho - \rho_m$;

$0 < C < \infty$: $n_{k,1,1,\ldots,1} = (\beta\pi^2(k^2 - 1) + E)^{-1}$, $k = 1,2,\ldots$, where E is the positive solution of $\sum_{k=1}^{\infty} (\beta\pi^2(k^2 - 1) + E)^{-1} = C^{-1}(\rho - \rho_m)$;

$C = \infty$: none of the single-particle states will be macroscopically occupied.

Proof: Suppose $\rho > \rho_c$. From the corollary of Sec. III we see $d/2 + 1/\alpha_1 + \cdots + 1/\alpha_d > 1$. Fix $\varepsilon > 0$ and take λ_ε as in (11). Consider the expression

$$\sum_{\lambda_k > m_d L_d^{-\gamma_d}} n_k(L_1,\ldots,L_d) = \int_{[m_d L_d^{-\gamma_d}, \lambda_\varepsilon)} \frac{z(L_1,\ldots,L_d)}{e^{\beta\lambda} - z(L_1,\ldots,L_d)}\, dF_{L_1,\ldots,L_d}(\lambda) + \int_{[\lambda_\varepsilon, \infty)} \frac{z(L_1,\ldots,L_d)}{e^{\beta\lambda} - z(L_1,\ldots,L_d)}\, dF_{L_1,\ldots,L_d}(\lambda). \tag{19}$$

To prove the existence of the limit ($L_d \to \infty$) we observe that the second expression in the rhs simply converges to

$$\int_{[\lambda_\varepsilon, \infty)} \frac{1}{e^{\beta\lambda} - 1}\, dF(\lambda)$$

by (3) and the fact that $z/(e^{\beta\lambda} - z) \leqslant 1/(e^{\beta\lambda} - 1)$, $z \leqslant 1$, the latter being bounded on $[\lambda_\varepsilon, \infty)$. From (18) and Lemma 2 we conclude

$$F_{L_1,\ldots,L_d}(\lambda) \leqslant C_1(\lambda + dC_2 L_d^{-\gamma_d})^{d/2 + 1/\alpha_1 + \cdots + 1/\alpha_d} \leqslant C_1[(1 + (C_2/m_d)d)\lambda]^{d/2 + 1/\alpha_1 \cdots + 1/\alpha_d}$$

for all λ in $[m_d L_d^{-\gamma_d}, \lambda_\varepsilon)$. It follows that $\text{const} \cdot 1/(e^{\beta\lambda} - 1) \cdot \lambda^{d/2 + 1/\alpha_1 + \cdots + 1/\alpha_d}$ is a integrable (on $[0, \lambda_\varepsilon)$) majorant of the integrand of the first term of (19). By the dominated convergence principle we get

$$\lim_{L_d \to \infty} \sum_{\lambda_k > m_d L_d^{-\gamma_d}} n_k(L_1,\ldots,L_d) = \int_{[0,\infty)} \frac{1}{e^{\beta\lambda} - 1}\, dF(\lambda) = \rho_c. \tag{20}$$

The relations (20) and (7) together yield

$$\rho - \rho_c = \lim_{L_d \to \infty} \sum_{k = (k_1,\ldots,k_{d-1})} n_k(L_1,\ldots,L_d) = \lim_{L_d \to \infty} \frac{1}{L_d} \int_{[0,\infty)} \frac{z(L_1,\ldots,L_d)}{e^{\beta\lambda} - z(L_1,\ldots,L_d)}\, dF_{L_1,\ldots,L_{d-1}}(\lambda).$$

The method now becomes clear. One restarts from (19) with $d - 1$ and repeats the conclusions made in the discussion of relation (20). Let us do so now.

To shorten the notation denote $s = d - 1$. Again we split the sum into two parts:

$$\sum_{\substack{\lambda_k > m_s L_s^{-\gamma_s} \\ k = (k_1,\ldots,k_s,1)}} n_k(L_1,\ldots,L_d) = \frac{1}{L_d}\left(\int_{[m_s L_s^{-\gamma_s}, \lambda_\varepsilon)} + \int_{[\lambda_\varepsilon, \infty)} \frac{z(L_1,\ldots,L_d)}{e^{\beta\lambda} - z(L_1,\ldots,L_d)} \times dF_{L_1,\ldots,L_d}\right). \tag{21}$$

We are interested in whether the term in parentheses is bounded. Again only the first of the integrals is critical. Re-

peating the arguments following (19) we arrive at a majorant $[1/(e^{\beta\lambda}-1)]\,\lambda^{s/2+1/\alpha_1+\cdots+1/\alpha_s}$ which is integrable at zero if and only if $s/2+1/\alpha_1+\cdots+1/\alpha_\gamma>1$. This will be the case for $s\geqslant 2$ if $1/\alpha_1+1/\alpha_2>0$ and for $s\geqslant 3$ if $1/\alpha_1+1/\alpha_2=0$ and consequently (21) tends to zero in all these cases. Iterating now $s=d-1,d-2,\ldots$ one concludes from (20) and the convergence to zero in the cases described above of (21) that

$$\rho-\rho_c=\lim_{L_d\to\infty}\sum_{k=(k_1,1,\ldots,1)} n_k(L_1,\ldots,L_d),\quad \text{if}\ \frac{1}{\alpha_1}+\frac{1}{\alpha_2}>0 \tag{22}$$

and

$$\rho-\rho_c=\lim_{L_d\to\infty}\sum_{k=(k_1,k_2,1,\ldots,1)} n_k(L_1,\ldots,L_d),$$
$$\text{if}\ \frac{1}{\alpha_1}+\frac{1}{\alpha_2}=0. \tag{23}$$

(i)Case $1/\alpha_1+1/\alpha_2>0$

Denote $\alpha=\alpha_1, m=m_1, \gamma=\gamma_1$, and $L=(L_1,\ldots,L_d)$. From (22) we have

$$\rho-\rho_c=\lim n_{1,1,\ldots,1}(L)+S(L), \tag{24a}$$

where

$$S(L)=\sum_{k=2}^{\infty} n_{k,1,1,\ldots,1}(L), \tag{24b}$$

$$S(L)=\frac{1}{L_2\cdot\cdots\cdot L_d}\left(\int_{[mL_1^{-\gamma},\lambda_\varepsilon)}+\int_{[\lambda_\varepsilon,\infty)}\frac{z(L_1,\ldots,L_d)}{e^{\beta\lambda}-z(L_1,\ldots,L_d)}\times dF_{L_1}(\lambda)\right). \tag{25}$$

$\alpha<2$: deduces from $F_{L_1}(\lambda)\leqslant C_1(\lambda+C_2L_1^{-\gamma})^{1/2+1/\alpha}$ the existence of a majorant $[1/(e^{\beta\lambda}-1)]\lambda^{1/2+1/\alpha}$ of the integrand in question which is integrable at zero because $\frac{1}{2}+\frac{1}{\alpha}>1$. Therefore, $S(L)$ tends to zero.

$\alpha=2$: Define $\omega_L=1-z(L)$, ω_L tending to zero if $\rho>\rho_c$. In Appendix B we will show the following two relations:

$$\limsup_{L\to\infty} S(L)\leqslant B/2\beta\sqrt{a_1}=\rho_m-\rho_c, \tag{26}$$

and if $\{\omega_L\cdot L_1\}$ is bounded then

$$\liminf_{L\to\infty} S(L)\geqslant B/2\beta\sqrt{a_1}=\rho_m-\rho_c. \tag{27}$$

If $\rho_c<\rho<\rho_m$ then $\lim_{L\to\infty}\omega_L\cdot L_1=\infty$, because from (24) one always has $S(L)\leqslant\rho-\rho_c<\rho_m-\rho_c$. Therefore, relation (27) cannot be valid. This implies

$$\lim_{L\to\infty} n_{1,1,\ldots,1}(L)=\lim_{L\to\infty}\frac{1-\omega_L}{L_1\cdot\cdots\cdot L_d\cdot\omega_L}=0,$$

and consequently none of the single-particle states are macroscopically occupied.

If $\rho>\rho_m$ we have from (24) and (26)

$$\liminf_{L\to\infty} n_{1,1,\ldots,1}(L)=\liminf_{L\to\infty}(\rho-\rho_c-S(L))$$
$$\geqslant\rho-\rho_c-(\rho_m-\rho_c)$$
$$\geqslant\rho-\rho_m>0.$$

It follows that $\omega_L<((\rho-\rho_m)\cdot L_1\cdot\cdots\cdot L_d)^{-1}$, and $\omega_L\cdot L_1$ tends to zero. Now relation (27) holds and consequently

$$\lim_{L\to\infty} n_{1,1,\ldots,1}(L)=n_{1,1,\ldots,1}=\rho-\rho_m.$$

Moreover the ground state is the unique state which is macroscopically occupied because from $\lambda_k>mL_1^{-1}$ for all $k\neq(1,1,\ldots,1)$ we see that

$$\lim_{L\to\infty} n_k(L)\leqslant\lim_{L\to\infty}\frac{1}{L_1\cdot\cdots\cdot L_d}\cdot\frac{1}{mL_1^{-1}+\omega_L}=0.$$

$\alpha>2$: Again we are interested in the first part of (25) because the second is bounded and therefore multiplication by $(L_2\cdot\cdots\cdot L_d)^{-1}$ yields a term tending to zero as $L\to\infty$. We shall use partial integration[8] and get

$$\int_{[\varepsilon,\infty)}\frac{z}{e^{\beta\lambda}-z}\,dF_L(\lambda)$$
$$=\beta\int_{[\varepsilon,\infty)}\frac{ze^{\beta\lambda}}{(e^{\beta\lambda}-z)^2}(F_L(\lambda)-F_L(\varepsilon))d\lambda. \tag{28}$$

Consider

$$\beta\int_{[mL_1^{-\gamma},\lambda_\varepsilon)}\frac{ze^{\beta\lambda}}{(e^{\beta\lambda}-z)^2}\left(F_{L_1}(\lambda)-\frac{1}{L_1}\right)d\lambda$$
$$\leqslant\frac{e^{\beta\lambda_\varepsilon}}{\beta}\cdot C_2\left(1+\frac{C_1}{m}\right)^{1/2+1/\alpha}\int_{[mL_1^{-\gamma},\lambda_\varepsilon)}\frac{\lambda^{1/2+1/\alpha}}{(\lambda+\beta^{-1}\omega_L)^2}\,d\lambda$$
$$\leqslant\text{const}\int_{[mL_1^{-\gamma},\lambda_\varepsilon)}(\lambda+\beta^{-1}\omega_L)^{-3/2+1/\alpha}\,d\lambda$$
$$=\text{const}\left(\frac{\alpha-2}{2\alpha}(mL^{-\gamma}+\beta^{-1}\omega_L)^{-(\alpha-2)/2\alpha}-\frac{\alpha-2}{2\alpha}(\lambda_\varepsilon+\beta^{-1}\omega_L)^{-(\alpha-2)/2\alpha}\right).$$

We again neglect the second term because it is bounded. The first is equal to

$$\text{const}\,L^{(\alpha-2)/(\alpha+2)}(m+\beta^{-1}\omega_L\cdot L^{\gamma})^{-(\alpha-2)/2\alpha}. \tag{29}$$

If now $D=0$ then $\lim S(L)=0$, and consequently none but the ground state is macroscopically occupied.

$0<D<\infty$ will imply $\{\omega_L\cdot L_1^{\gamma}\}$ is bounded. Suppose to the contrary $\lim_{L\to\infty}\omega_L L_1^{\gamma}=\infty$. Then $\lim S(L)=0$ from (29). But

$$\lim n_{1,1,\ldots,1}(L)=\lim_{L\to\infty}\frac{1}{L_1\cdot\cdots\cdot L_d\cdot\omega_L}$$
$$=\lim_{L\to\infty}\frac{L_1^{(\alpha-2)/(\alpha+2)}}{L_2\cdot\cdots\cdot L_d\cdot\omega_L\cdot L_1^{\gamma}}=0$$

in contrast to (24) because $\rho-\rho_c<0$.

From relation (6) we have $\lambda_k\leqslant M_kL_1^{-\gamma}$ which yields a lower bound for the occupation number

$$n_{k,1,1,\ldots,1}(L)\geqslant\frac{(1-\omega_L)(1-\lambda_k)}{L_1\cdot\cdots\cdot L_d\cdot(\beta M_kL_1^{-\gamma}+\omega_L(1-\lambda_k))}, \tag{30}$$

where we used the inequality $e^x-1\leqslant x/(1-x)$ for $x<2$. As $\{\omega_L\cdot L_1^{\gamma}\}$ is bounded we get a lower bound from (30)

$$\liminf_{L\to\infty} n_{k,1,1,\ldots,1}(L)$$

$$\geqslant \liminf_{L\to\infty} \frac{L^{(\alpha-2)/(\alpha+2)}}{L_2\cdot\cdots\cdot L_d}\cdot\frac{1}{\beta M_k+\omega_L L_1^{\gamma}}>0.$$

$D=\infty$: From

$$\lim_{L\to\infty} n_{1,1,\ldots,1}(L)=\lim_{L\to\infty}\frac{1}{L_1\cdot\cdots\cdot L_d\cdot\omega_L}\leqslant\rho-\rho_c<\infty$$

we conclude

$$\infty>\lim_{L\to\infty}\frac{1}{L_1\cdot\cdots\cdot L_d\cdot\omega_L}=\lim_{L\to\infty}\frac{L_1^{(\alpha-2)/(\alpha+2)}}{L_2\cdot\cdots\cdot L_d}\cdot\frac{1}{\omega_L\cdot L_1^{\gamma}}$$

and finally $\lim \omega_L L_1^{\gamma}=\infty$ as $D=\infty$.

Again, from (30) we conclude

$$\liminf_{L\to\infty} n_{k,1,1,\ldots,1}(L)$$

$$\geqslant\lim_{L\to\infty}\frac{1}{L_1\cdot\cdots\cdot L_d(\beta M_k L_1^{-\gamma}+\omega_L)}$$

$$\geqslant\lim_{L\to\infty}\frac{1}{L_1\cdot\cdots\cdot L_d\cdot\omega_L(\beta M_k(\omega_L\cdot L_1^{\gamma})^{-1}+1)}$$

$$=n_{1,1,\ldots,1}.$$

Consequently we have a equipartition of the condensate. Thus $n_k=0$ for all $k=(k_1,k_2,\ldots,k_d)$.

(ii)Case $1/\alpha_1+1/\alpha_2=0$

This corresponds to Dirichlet boundary conditions and is actually identical with a result of Van den Berg.[3] There is some discrepancy concerning ρ_m, coming from the fact that we handle here a prism of sides $2\times L_i$ instead of L_i and the fact that we have $-\Delta$ instead of $-\Delta/2$. Our proof is actually much more simpler and intuitive. Equation (23) proves one half of the theorem. We sketch the remainder. From (23) we have

$$\rho-\rho_c=\lim_{L\to\infty}\Big(\sum_{k=(k_1,1,\ldots,1)} n_k(L)+S(L)\Big),\tag{31}$$

where

$$S(L)=\sum_{\substack{k=(k_1,k_2,1,1,\ldots,1)\\ k_2>2}} n_k(L)\tag{32}$$

$$=\frac{1}{L_3\cdot\cdots\cdot L_d}\int_{[\frac{3}{4}\pi^2 L_2^{-2},\infty)}\times\frac{z(L)}{e^{\beta\lambda}-z(L)}\,dF_{L_1,L_2}(\lambda).\tag{33}$$

To derive the limit $L\to\infty$ we need estimates on $F_{L_1,L_2}(\lambda)$. But for Dirichlet eigenvalues we have simply[8,10]

$$-\frac{\text{const}}{L_1\cdot L_2}(1+\sqrt{\lambda}\,(L_1+L_2))+\frac{\lambda}{\pi}$$

$$\leqslant F_{L_1,L_2}(\lambda)\leqslant\frac{\lambda^{+}}{\pi}+\frac{\text{const}}{L_1\cdot L_2}(1+\sqrt{\lambda^{+}}\,(L_1+L_2))$$

where

$$\lambda^{+}=\lambda+E_{1,1,\ldots,}(L).$$

From this one obtains in a similar way as in Appendix B

$$\lim_{L\to\infty} S(L)=2A/\pi\beta=\rho_m-\rho_c.$$

ACKNOWLEDGMENTS

The author would like to thank Dr. M. Schröder for helpful discussions. Hints and valuable remarks of the referee are gratefully acknowledged.

APPENDIX A

To calculate a lower bound for the distances of the eigenvalues of an operator $-d^2/dx^2+V(x)$ $(d=1)$ with eigenvalues E_n we cite here a result of Kirsch and Simon.[13] Define

$$\lambda=\max_{\substack{E\in[E_{n-1},E_n]\\ x\in(a,b)}}|E-V(x)|^{1/2}.\tag{A1}$$

They show: If V is continuous (but perhaps not bounded above), and suppose that for $A>0$, $V(x)\geqslant E_n+A^2$ on $\mathbb{R}\setminus[u,v]$, where $E_n>E_{n-1}$ are eigenvalues of $-d^2/dx^2+V(x)$. Then

$$E_n-E_{n-1}\geqslant\pi\cdot\lambda^2\cdot A\cdot(\lambda+A)^{-1}\exp[-\lambda(v-u)].\tag{A2}$$

We want to use this for our operator $H_L=-d^2/dx^2+v(x/L)$ with eigenvalues $E_1(L)<E_2(L)\leqslant\cdots$. From (6) we have a constant $M, E_2(L)\leqslant M\cdot L^{-2\alpha/(\alpha+2)}$, take $A=L^{-\alpha/(\alpha+2)}$, and a symmetric interval $(-b,b)$ with $b=BL^{\alpha/(\alpha+2)}$, where

$$B=\max\left(\left(\frac{M+1}{a-\varepsilon}\right)^{1/\alpha},1\right)\quad 0<\varepsilon<a.$$

Take L large enough to secure $B\cdot L^{-2/(\alpha+2)}<x_\varepsilon$. If $|x|\geqslant b$ then

$$V(x/L)\geqslant(a-\varepsilon)|BL^{-2/(\alpha+2)}|^{\alpha}$$

$$\geqslant(M+1)\cdot L^{-2\alpha/(\alpha+2)}\geqslant E_2(L)+A.\tag{A3}$$

If $|x|<b$ then

$$V\left(\frac{x}{L}\right)\leqslant(a+\varepsilon)|BL^{-2/(\alpha+2)}|^{\alpha}$$

$$\leqslant\max\left(\frac{a+\varepsilon}{a-\varepsilon}(M+1),a+\varepsilon\right)L^{-2\alpha/(\alpha+2)}.\tag{A4}$$

One concludes

$$\lambda\leqslant\lambda^{+}=\max\left(\sqrt{M},\sqrt{\max\left(\frac{a+\varepsilon}{a-\varepsilon}(M+1),a+\varepsilon\right)}\right)$$

$$\times L^{-\alpha/(\alpha+2)}=Q\times L^{-\alpha/(\alpha+2)}$$

$$\lambda\geqslant\lambda^{-}=A.$$

The function

$$f(\lambda)=\lambda^2 A\cdot(\lambda+A)^{-1}e^{-2\lambda b},\quad\lambda\geqslant 0$$

reaches its maximum at

$$\lambda_0=(A/4B)(\sqrt{4B^2+12B+1}-2B+1).$$

We have $\lambda_0<A$ if $B\geqslant 1$ such that we have $\lambda_0\leqslant\lambda^{-}\leqslant\lambda\leqslant\lambda^{+}$. From this and from (A2) we see

$$E_2(L)-E_1(L)\geqslant\pi\lambda^2 A\cdot(\lambda+A)^{-1}e^{-2\cdot\lambda\cdot b}$$

$$\geqslant\pi(\lambda^{+})^2 A\cdot(\lambda^{+}+A)^{-1}e^{-2\cdot\lambda^{+}b}$$

$$\geqslant\pi Q^2(Q+1)^{-1}e^{-2\cdot Q\cdot B}\cdot L^{-2\alpha/(\alpha+2)}.\tag{A5}$$

Now the proof of (7) is complete.

APPENDIX B

From Lemma 1 we have for $\alpha = 2$

$$-\frac{N}{L_1}+\frac{A_N}{\pi(a+\varepsilon)^{1/2}}\lambda \leqslant F_{L_1}(\lambda)\leqslant \frac{B_N}{\pi(a-\varepsilon)^{1/2}}(\lambda + M_1 L_1^{-1}) + \frac{N+2}{L_1}. \quad \text{(B1)}$$

We use throughout the appendix the partial integrated form (28) of our integrals. Note that

$$\limsup_{L\to\infty} S(L) = \limsup_{L\to\infty} \frac{1}{L_2\cdot\cdots\cdot L_d}\beta \times \int_{[mL_1^{-1},\lambda_\varepsilon)} \frac{z(L)e^{\beta\lambda}}{(e^{\beta\lambda}-z(L))^2} F_{L_1}(\lambda)\,d\lambda. \quad \text{(B2)}$$

In order to use (B1) we first remark

$$\frac{1}{L_1\cdot L_2\cdot\cdots\cdot L_d}\int_{[mL_1^{-1},\lambda_\varepsilon)}\frac{d\lambda}{\lambda^2} = \frac{1}{L_1\cdot\cdots\cdot L_d}\left(\frac{L_1}{m}-\lambda_\varepsilon\right). \quad \text{(B3)}$$

Secondly we use the following estimate for the integrand of (B2):

$$\frac{z(L)e^{\beta\lambda}}{(e^{\beta\lambda}-z(L))^2}F_{L_1}(\lambda) \leqslant \frac{e^{\beta\lambda}}{(\beta\lambda)^2}\left(\frac{B_N}{\pi(a-\varepsilon)^{1/2}}\lambda+\frac{C}{L_1}\right). \quad \text{(B4)}$$

Keeping in mind relations (B1)–(B4) we conclude

$$\begin{aligned}\limsup_{L\to\infty} S(L) &\leqslant \lim_{L\to\infty}\frac{1}{L_2\cdot\cdots\cdot L_d}\frac{B_N}{\beta\pi(a-\varepsilon)^{1/2}}\int_{[mL_1^{-1},\lambda_\varepsilon)}\frac{e^{\beta\lambda}\,d\lambda}{\lambda}\\ &= \lim_{L\to\infty}\frac{1}{L_2\cdot\cdots\cdot L_d}\frac{B_N}{\beta\pi(a-\varepsilon)^{1/2}} \times(\log\lambda_\varepsilon - \log mL_1^{-1})\\ &= \frac{B\cdot B_N}{\beta\pi(a-\varepsilon)^{1/2}}. \end{aligned} \quad \text{(B5)}$$

In (B5) we simply ignored $e^{\beta\lambda}$ which is correct in the limit. Integrating term by term $1/\lambda + 1 + \lambda/2! + \cdots$ all but the first give zero in the limit.

To calculate a lower bound we have

$$\frac{z(L)e^{\beta\lambda}}{(e^{\beta\lambda}-z(L))^2}F_{L_1}(\lambda) \geqslant \frac{1}{\beta^2}\frac{(1-\omega_L)(1-\lambda_\varepsilon\beta)^2}{(\lambda+(\beta^{-1}-\lambda_\varepsilon)\omega_L)^2} \times\left(\frac{A_N}{\pi(a+\varepsilon)^{1/2}}\lambda-\frac{N}{L_1}\right) \quad \text{(B6)}$$

if $\beta\lambda_\varepsilon < 1$. This will be the case if ε is small enough. For $\lambda < \lambda_\varepsilon$ one has $e^{\beta\lambda}-1\leqslant\beta\lambda/(1-\beta\lambda_\varepsilon)$ such that (B6) follows. From (B1), (B2), (B3), and (B6) one concludes now

$$\begin{aligned}\liminf_{L\to\infty} S(L) &\geqslant \frac{A_N(1-\lambda_\varepsilon\beta)^2}{\beta\pi(a+\varepsilon)^{1/2}}\cdot\liminf_{L\to\infty}\frac{1-\omega_L}{L_2\cdot\cdots\cdot L_d}\\ &\times\int_{[mL_1^{-1},\lambda_\varepsilon)}\frac{d\lambda}{(\lambda+(\beta^{-1}-\lambda_\varepsilon)\omega_L)^2}. \end{aligned} \quad \text{(B7)}$$

The integral in (B7) can easily be calculated

$$\begin{aligned}\int\frac{d\lambda}{(\lambda+(\beta^{-1}-\lambda_\varepsilon)\omega_L)^2} &= \log(\lambda+(\beta^{-1}-\lambda_\varepsilon)\omega_L)\\ &+(\beta^{-1}-\lambda_\varepsilon)\omega_L\cdot(\lambda+(\beta^{-1}-\lambda_\varepsilon)\omega_L)^{-1}. \end{aligned} \quad \text{(B8)}$$

Again only the value at $m\cdot L_1^{-1}$ will be important,

$$-\log L_1 + (\log(m+(\beta^{-1}-\lambda_\varepsilon)\omega_L\cdot L_1) + \left(\frac{(\beta^{-1}-\lambda_\varepsilon)}{m(\omega_L\cdot L_1)^{-1}+(\beta^{-1}-\lambda_\varepsilon)}\right). \quad \text{(B9)}$$

Clearly the terms in parentheses are bounded if $\omega_L\cdot L_1 < \infty$. Claiming this one gets finally

$$\liminf S(L) \geqslant \frac{B\cdot A_N(1-\lambda_\varepsilon\beta)^2}{\beta\pi(a+\varepsilon)^{1/2}}. \quad \text{(B10)}$$

The relations (26) and (27) follow now by taking the limits $N\to\infty$, and $\varepsilon\to 0$, because $\lim_{\varepsilon\to 0}\lambda_\varepsilon = 0$, and $\lim A_N = \lim B_N = \pi/2$ from (8b).

[1] M. Van den Berg, J. T. Lewis, and J. V. Pulè, Helv. Phys. Acta **59**, 1271 (1986).
[2] M. Van den Berg, J. Math. Phys. **23**, 1159 (1982).
[3] M. Van den Berg, J. Stat. Phys. **31**, 623 (1983).
[4] M. Van den Berg and J. T. Lewis, Physica **110A**, 550 (1982).
[5] M. Van den Berg and J. T. Lewis, Commun. Math. Phys. **81**, 475 (1981).
[6] H. B. G. Casimir, *On Bose-Einstein Condensation. Fundamental Problems in Statistical Mechanics II*, edited by E. G. D. Cohen (North-Holland, Amsterdam, 1968), pp. 188–196.
[7] To secure this choose, e.g., $V_i(x_i) > a_i|x_i|^{\alpha_i}$, $i = 1,2,\ldots,d$.
[8] M. Reed and B. Simon, *Methods of Modern Mathematical Physics, Analysis of Operators* (Academic, New York, 1978), Vol. IV.
[9] In fact, Pulè proves his Lemma 1 for $-(\Delta/2)+V(x/L)$ on an open bounded region. Λ^L with Dirichlet conditions on $\partial\Lambda^L$. The necessary generalization consists now in making the scaling factors $L_1,L_2,\ldots,L_d$ of the axes independent of each other. Λ^L will be replaced by $\Lambda^{L_1,L_2,\ldots,L_d}$ and $V(x/L)$ by $V(x_1/L_1,x_2/L_2,\ldots,x_d/L_d)$. The difference between this generalization and our "boundary free" Lemma 3 is not essential. In the proof of our lemma one has to substitute the set $\Lambda^{L_1,L_2,\ldots,L_d}$ by the set $\{x\in\mathbb{R}^d: V(x_1/L_1,x_2/L_2,\ldots,x_d/L_d)<\lambda\}$.
[10] J. V. Pulè, J. Math. Phys. **24**, 138 (1983).
[11] K. Voigt, Wiss. Z. KMU Leipzig MNR **34**, 651 (1985).
[12] The formulation in Sec. I of our problem seems to contain a uncertainty concerning the volume of the system. The divisor $L_1\cdot L_2\cdot\cdots\cdot L_d$ in the definition of $F_{L_1,L_2,\ldots,L_d}$ plays the role of the volume of the gas but its setting is completely arbitrary. Remark first that our assumption that $V(x)\to\infty$ as $x\to\infty$ "localizes" the gas near the origin. This becomes clear in the more extreme case $\alpha_1=\alpha_2=\cdots\alpha_d=\infty$, where the gas gets enclosed in a rectangular parallelepiped with volume $2^d L_1\cdot\cdots\cdot L_d$. Secondly our results remain in principal unchanged in the case of a boson gas enclosed in a rectangular parallelepiped in the presence of an external potential $V(x_1/L_1,x_2/L_2,\ldots,x_d/L_d)$ where our divisor in question describes the correct volume of the gas.
[13] W. Kirsch and B. Simon, Commun. Math. Phys. **97**, 453 (1985).

Completely solvable models of the nonlinear Boltzmann equation I. Case of three velocities

Per Osland
Department of Physics, University of Bergen, Allégt. 55, N-5007 Bergen, Norway

Tai Tsun Wu
Gordon McKay Laboratory, Harvard University, Cambridge, Massachusetts 02138

(Received 29 May 1990; accepted for publication 8 August 1990)

In one space and one time dimension, a class of models of the nonlinear Boltzmann equation is presented that is exactly solvable for all initial conditions. The models have three velocity components and the following desirable properties: (a) conservation of the number of particles; (b) energy conservation; (c) nonlinearity; (d) positivity of distribution functions; and (e) unique equilibrium state (for any given density), which is approached as $t\to\infty$. These models are very rich in structure, and some of their simple properties are studied.

I. INTRODUCTION

Because of the inherent complexity of the nonlinear Boltzmann transport equation, a great deal of insight has been provided by simplified models where some exact solutions can be found explicitly. For example, the exact solution discovered by Krook and Wu,[1] and also by Bobylev,[2] in connection with controlled thermonuclear fusion, has been used by Olaussen[3] to destroy a conjecture[4] concerning the approach to equilibrium. There is by now a rather large number of such particular exact solutions[5] to various models of the nonlinear Boltzmann equation, including an especially interesting class given by Cornille.[6]

These particular exact solutions are all of the following nature: they are exact solutions to well-defined models of the Boltzmann equation and contain a finite number of arbitrary constants. In none of these cases can these exact solutions be used to solve the general initial problem, because such an initial value problem necessarily includes one or more arbitrary functions giving the initial distributions. It is for this reason that these solutions are referred to as particular exact solutions.

It is clearly desirable to have some simplified model of the Boltzmann equation where the exact solution can be written down for arbitrary initial conditions. It was nearly a decade ago when the first such model was found.[7] The input that made that attempt successful came from particle physics. Christenson *et al.*[8] found experimentally a quarter of a century ago that time reversal is not an invariance for weak interactions. Even before this beautiful and decisive experiment, it was known that time reversal invariance does not play a central role in kinetic theory.[9] Accordingly, in a model for the nonlinear Boltzmann equation, not only is there no compelling reason, but indeed it may not even be desirable, to satisfy detailed balance.

With this point in mind, in the proposed model[7] there is one space dimension and of course one time dimension, and there are only two allowed velocities, called $+1$ and -1, with corresponding distribution functions $f_1(x,t)$ and $f_2(x,t)$, respectively. The model Boltzmann equation consists of

$$(\partial_t+\partial_x)f_1=f_1f_2-\alpha f_1+\beta f_2, \tag{1.1a}$$

$$(\partial_t-\partial_x)f_2=-f_1f_2+\alpha f_1-\beta f_2, \tag{1.1b}$$

where α and β are two positive constants. It is seen that there are three collision processes as follows:

$$+\,-\to+\,+\qquad 1, \tag{1.2}$$

$$+\to-\qquad \alpha, \tag{1.3}$$

$$-\to+\qquad \beta. \tag{1.4}$$

The second and the third collision processes may be interpreted as collisions with some kind of background gas.

Ever since the discovery of this completely solvable model, we have been making a continual effort to look for other models of this nature. The first hint of the existence of larger classes of completely solvable models came in 1985. It is the purpose of this paper to describe the next simplest such model that has been found so far, and to give some of the more elementary properties.

II. BASIC IDEA

In the model given previously,[7] there are, as already mentioned, two allowed velocities. In the present paper, we consider the case of three allowed velocities. However, this increase from two velocities to three velocities turns out to be most difficult. The reason is as follows. In Ref. 7, the way to solve the model Boltzmann equation is to express both of the distribution functions f_1 and f_2 as suitable derivatives of a single function, called F. That the model is completely solvable comes from the fact that the exponential of F satisfies a linear partial differential equation. Since there are only two independent variables, namely x and t, this F can have only two independent derivatives. Accordingly, how is it possible to have more than two allowed velocities?

The central point here is to avoid introducing F altogether. In other words, it is essential to work on the level of the f's rather than on the level of the F, even though in the

 0022-2488/90/123003-27$03.00

previous, much simpler case it was the introduction of F that led to the complete solution.

The model we propose to study has three distribution functions, $f_1(x,t)$, $f_2(x,t)$, and $f_3(x,t)$, associated with the velocities v_1, v_2, and v_3, respectively. These distribution functions satisfy the following nonlinear Boltzmann equation:

$$\begin{aligned}(\partial_t + v_1\partial_x)f_1 &= f_1[a_3f_2 - a_2f_3] - (\alpha+\alpha')f_1 + \beta f_2 \\ &\quad + \gamma' f_3, \\ (\partial_t + v_2\partial_x)f_2 &= f_2[a_1f_3 - a_3f_1] - (\beta+\beta')f_2 + \gamma f_3 \\ &\quad + \alpha' f_1, \\ (\partial_t + v_3\partial_x)f_3 &= f_3[a_2f_1 - a_1f_2] - (\gamma+\gamma')f_3 + \alpha f_1 \\ &\quad + \beta' f_2,\end{aligned} \tag{2.1}$$

where the a_i and $\alpha,\beta,\ldots,\gamma'$ are constants. The latter are non-negative,

$$\alpha \geqslant 0, \quad \beta \geqslant 0, \ldots, \quad \gamma' \geqslant 0. \tag{2.2}$$

The model is constructed such that the total number of particles is conserved, i.e., the distribution functions obey the continuity equation,

$$\partial_t \sum_{i=1}^{3} f_i + \sum_{i=1}^{3} v_i \partial_x f_i = 0. \tag{2.3}$$

Of course, the three distribution functions are required to be non-negative,

$$f_i \equiv f_i(x,t) \geqslant 0, \quad i=1,2,3. \tag{2.4}$$

The collision terms $\pm\, a_3 f_1 f_2$ represent collisions in which (with $a_3 > 0$) particles of initial velocities v_1 and v_2 undergo an interaction and both achieve velocity v_1. The linear collision terms represent interactions with some kind of background gas: The terms $\pm\, \alpha f_1$ represent particles of species 1 interacting with the background and turning into particles of species 3. The different collision processes are summarized in Table I.

We are primarily interested in the case when all three velocities are different. It is clear from Table I that the number of particles is conserved. In order for energy to be conserved, the particles having different velocities should also have different masses and/or different internal energies. In kinetic theory, the conservation of momentum is usually not important. However, if it is desired to have also momentum conservation, it is only necessary to choose the three velocities to have the same sign. In this case, let m_1, m_2, and m_3 be the masses of the particles with the velocities v_1, v_2, and v_3, respectively, and let e_1, e_2, and e_3 be the corresponding internal or excitation energies. Then both energy and momentum are conserved if, with non-relativistic kinematics,

$$m_1v_1 = m_2v_2 = m_3v_3, \tag{2.5a}$$

and

$$\tfrac{1}{2}m_1v_1^2 + e_1 = \tfrac{1}{2}m_2v_2^2 + e_2 = \tfrac{1}{2}m_3v_3^2 + e_3. \tag{2.5b}$$

TABLE I. Collision processes.

Collision process	Relative probability
$a_1>0, \quad a_2>0, \quad a_3>0$:	
$v_2v_3 \to v_2v_2$	a_1
$v_3v_1 \to v_3v_3$	a_2
$v_1v_2 \to v_1v_1$	a_3
$a_1>0, \quad a_2<0, \quad a_3>0$:	
$v_2v_3 \to v_2v_2$	a_1
$v_3v_1 \to v_1v_1$	$\lvert a_2\rvert$
$v_1v_2 \to v_1v_1$	a_3
$v_1 \to v_2$	α'
$v_1 \to v_3$	α
$v_2 \to v_3$	β'
$v_2 \to v_1$	β
$v_3 \to v_1$	γ'
$v_3 \to v_2$	γ

When v_1, v_2, and v_3 have the same sign, these equations have a two-parameter family of solutions; furthermore, the three masses are all positive, as they should be.

In order to overcome the major problem discussed at the beginning of this section, the introduction of the function F is avoided here. Instead, the basic idea, which leads to solvable equations, is to take out a common denominator from the distribution functions. Let

$$f_i(x,t) = g_i(x,t)/D(x,t), \quad i=1,2,3. \tag{2.6}$$

Clearly, this $D(x,t)$ must be nonzero everywhere,

$$D(x,t) \neq 0, \quad \text{for all } x \text{ and all } t. \tag{2.7}$$

Substituting this form into Eq. (2.1), we obtain equations where *all terms are bilinear* in the four new functions g_1, g_2, g_3, and D:

$$\begin{aligned}&-g_1(\partial_t + v_1\partial_x)D + D(\partial_t + v_1\partial_x)g_1 \\ &\quad = g_1(a_3g_2 - a_2g_3) - (\alpha+\alpha')g_1D + D(\beta g_2 + \gamma' g_3), \\ &-g_2(\partial_t + v_2\partial_x)D + D(\partial_t + v_2\partial_x)g_2 \\ &\quad = g_2(a_1g_3 - a_3g_1) - (\beta+\beta')g_2D + D(\gamma g_3 + \alpha' g_1), \\ &-g_3(\partial_t + v_3\partial_x)D + D(\partial_t + v_3\partial_x)g_3 \\ &\quad = g_3(a_2g_1 - a_1g_2) - (\gamma+\gamma')g_3D + D(\alpha g_1 + \beta' g_2).\end{aligned} \tag{2.8}$$

We may now satisfy (2.8) by requiring simultaneously

$$\begin{aligned}-(\partial_t + v_1\partial_x)D &= a_3g_2 - a_2g_3 - \bar{\alpha}'D, \\ -(\partial_t + v_2\partial_x)D &= a_1g_3 - a_3g_1 - \bar{\beta}'D, \\ -(\partial_t + v_3\partial_x)D &= a_2g_1 - a_1g_2 - \bar{\gamma}'D\end{aligned} \tag{2.9}$$

and

$$(\partial_t + v_1\partial_x)g_1 = -\bar{\alpha}g_1 + \beta g_2 + \gamma' g_3,$$

$$(\partial_t + v_2\partial_x)g_2 = -\bar{\beta}g_2 + \gamma g_3 + \alpha' g_1, \tag{2.10}$$

$$(\partial_t + v_3\partial_x)g_3 = -\bar{\gamma}g_3 + \alpha g_1 + \beta' g_2.$$

The terms in (2.8) that are proportional to $g_i D$ have been split in an as yet unspecified way,

$$\begin{aligned} \bar{\alpha} + \bar{\alpha}' &= \alpha + \alpha', \\ \bar{\beta} + \bar{\beta}' &= \beta + \beta', \\ \bar{\gamma} + \bar{\gamma}' &= \gamma + \gamma'. \end{aligned} \tag{2.11}$$

Whereas all six parameters on the right-hand side of Eq. (2.11) are non-negative, we have no corresponding requirement to those on the left-hand side. Some might be negative.

It should be noted that the crucial feature of the model that leads to the linear equations (2.9)–(2.10) is the factor f_i in the quadratic collision term for the function f_i. Thus, in the equation for f_1, for example, there is no term $\sim f_2^2$. When we equate terms pairwise in Eq. (2.8), a factor g_i can thus be canceled, and one obtains (2.9).

It is seen that the approach here is quite different from that used previously in dealing with the two-velocity model.[7] This approach avoids the problem explained at the beginning of this section, and makes it possible to construct much larger and interesting classes of completely solvable models.

The four functions D, g_1, g_2, and g_3 must satisfy the six partial differential equations given by (2.9) and (2.10). While it is often possible to satisfy more equations than unknowns, what is needed here is much more. In order for this model to be completely solvable, i.e., exactly solvable for all initial conditions, we seek relations between the 12 parameters v_1, v_2, v_3, a_1, a_2, a_3, α, β, γ, α', β', and γ' such that these six equations are consistent for all initial conditions. Another way of stating our problem is as follows: For any initial conditions on f_1, f_2, and f_3, we look for relations between the 12 parameters such that, after solving (2.10) for g_1, g_2, and g_3, the three equations (2.9) are consistent so that a D can be found.

III. GENERALIZATIONS

The model can be generalized to more velocities. With N discrete velocities v_i and N distribution functions f_i, the model is given by

$$(\partial_t + v_i\partial_x)f_i = f_i \sum_{j\neq i} a_{ij}f_j + \sum_j \alpha_{ij}f_j, \quad i=1,\ldots,N, \tag{3.1}$$

with

$$\alpha_{ij} \geqslant 0 \text{ for } i \neq j. \tag{3.2}$$

Particles can switch from distribution f_j (with velocity v_j) to distribution f_i (with velocity v_i) at rates given by the coefficients of the nonlinear and linear collision terms, a_{ij} and α_{ij}. Conservation of the total number of particles,

$$\partial_t \sum_i f_i + \sum_i v_i\partial_x f_i = 0, \tag{3.3}$$

imposes the constraints

$$a_{ij} = -a_{ji} \tag{3.4}$$

and

$$\sum_i \alpha_{ij} = 0. \tag{3.5}$$

The basic idea described in the preceding section is applicable also in this case: There is one D and N g_j's, and these $N+1$ functions are required to satisfy $2N$ linear partial differential equations. Consistency can still be achieved by imposing suitable relations between the parameters.

One new feature appears when $N>3$ as distinct from the case $N=3$ of the present paper. Let $\mathscr{A}^{(N)}$ be the $N\times N$ matrix whose elements are the coefficients a_{ij}, then by (3.4) $\mathscr{A}^{(N)}$ is always antisymmetric. Since the rank of an antisymmetric matrix is always even,[10] the rank of $\mathscr{A}^{(3)}$ is 2. For $N>3$, the rank is not necessarily 2, and there may be interesting models where the rank of $\mathscr{A}^{(N)}$ is higher.

To a limited extent, these models can be generalized to the case of infinite N and also to two spatial dimensions. These generalizations to N larger than 3 and/or to two dimensions will be discussed in later publications.

IV. SCALING AND GALILEAN TRANSFORMATIONS

As already mentioned, there are 12 parameters in these models, namely, v_1, v_2, v_3, a_1, a_2, a_3, α, β, γ, α', β', and γ'. Scaling and Galilean transformations relate models with different values of parameters.

In all these transformations, v_1, v_2, and v_3 always transform the same way; so do the three a's. Similarly, the six parameters α, β, γ, α', β', and γ' always transform the same way. Therefore, in this section, it is convenient and more transparent to use the following short-hand notations:

$$\begin{aligned} &f \text{ stands for } f_1, f_2, \text{ or } f_3; \\ &v \text{ stands for } v_1, v_2, \text{ or } v_3; \\ &a \text{ stands for } a_1, a_2, \text{ or } a_3; \\ &\alpha \text{ stands for } \alpha, \beta, \gamma, \alpha', \beta', \text{ or } \gamma'. \end{aligned} \tag{4.1}$$

With this notation, we list explicitly the various scaling and Galilean transformations.

(a) First, the time variable may be scaled. Equations (2.1) are invariant under the following transformation:

$$\begin{aligned} &t \to c_t t, \\ &x \to x, \\ &f \to f, \\ &v \to v/c_t, \\ &a \to a/c_t, \end{aligned}$$

and

$$\alpha \to \alpha/c_t. \tag{4.2}$$

Because of (2.2), c_t must be positive.

(b) Second, the spatial variable may be similarly scaled. The corresponding transformation is

$$\begin{aligned} &t\to t,\\ &x\to c_x x,\\ &f\to f,\\ &v\to c_x v,\\ &a\to a,\\ &\alpha\to\alpha. \end{aligned} \tag{4.3}$$

Unlike c_t, this c_x can be positive or negative.

(c) Similarly, the distribution functions $f_j(x,t)$ may be scaled by a common factor c_f independent of both x and t. The corresponding transformation is

$$\begin{aligned} &t\to t,\\ &x\to x,\\ &f\to c_f f,\\ &v\to v,\\ &a\to a/c_f,\\ &\alpha\to\alpha. \end{aligned} \tag{4.4}$$

Because of (2.4), this c_f again must be positive.

(d) Finally, the Galilean transformation is

$$\begin{aligned} &t\to t,\\ &x\to x-Vt,\\ &f\to f,\\ &v\to v+V,\\ &a\to a,\\ &\alpha\to\alpha. \end{aligned} \tag{4.5}$$

Here, V can take on any real value.

These four transformations, three scaling and one Galilean, are independent. Therefore, of the 12 parameters, only eight are independent.

In two respects, the Galilean transformation (4.5) is of a slightly different nature from the scaling transformations (4.2)–(4.4). If energy conservation is required to hold after the Galilean transformation, then by (2.5b) the masses m_j and/or the internal energies e_j must also be suitably transformed. Second, in the discussion of possible equilibrium distributions in Sec. VI, (4.2)–(4.4) apply, but not (4.5), since equilibrium distributions that are not spatially homogeneous are not invariant under a Galilean transformation.

V. DESCRIPTION OF THE COMPLETELY SOLVABLE MODELS

In this section, we shall specify the subspace in the 12-dimensional parameter space where the nonlinear Boltzmann equation (2.1) can be explicitly and completely solved for all initial conditions. The conditions, Eq. (5.18) below, turn out to be two relations between the six parameters v_1, v_2, v_3, a_1, a_2, and a_3.

Throughout this section, it is assumed that the three v's are not all equal. In other words, the completely uninteresting case $v_1=v_2=v_3$ is excluded, because, by the Galilean transformation (4.5), all velocities can be reduced to zero and hence the Boltzmann equation (2.1) merely consists of three ordinary differential equations in the time variable. It is allowed to have two velocities equal, so long as they are not equal to the third.

In Sec. II we arrived at one set of partial differential equations involving g_1, g_2, and g_3 as well as D, Eq. (2.9), and another set, Eq. (2.10), involving g_1, g_2, and g_3 only. We shall consider the first set. If we multiply those three equations by a_1, a_2, and a_3, respectively, and add them, then the terms involving the g_i cancel:

$$\begin{aligned}[(a_1+a_2+a_3)\partial_t+(a_1v_1+a_2v_2+a_3v_3)\partial_x\\ -(a_1\bar\alpha'+a_2\bar\beta'+a_3\bar\gamma')]D=0.\end{aligned} \tag{5.1}$$

This first-order partial differential equation is either trivial, i.e.,

$$a_1+a_2+a_3=0, \tag{5.2}$$

$$a_1v_1+a_2v_2+a_3v_3=0, \tag{5.3}$$

$$a_1\bar\alpha'+a_2\bar\beta'+a_3\bar\gamma'=0, \tag{5.4}$$

or nontrivial, i.e., either (5.2) or (5.3) or both is not satisfied.

We want to show that the nonlinear Boltzmann equation (2.1) is completely solvable only if (5.1) is trivial; in other words, (5.2), (5.3), and (5.4) are all satisfied. In order to reach this conclusion, assume the opposite. With the notation

$$\begin{aligned} &A\equiv a_1+a_2+a_3,\\ &B\equiv a_1v_1+a_2v_2+a_3v_3,\\ &C\equiv a_1\bar\alpha'+a_2\bar\beta'+a_3\bar\gamma', \end{aligned} \tag{5.5}$$

(5.1) is

$$(A\partial_t+B\partial_x-C)D(x,t)=0, \tag{5.6}$$

and the most general solution is

$$D(x,t)=e^{pt+qx}D_0(Bt-Ax), \tag{5.7}$$

where the constants p and q satisfy

$$Ap+Bq-C=0, \tag{5.8}$$

and D_0 is a function of only *one* variable,

$$\xi\equiv Bt-Ax. \tag{5.9}$$

Define the following three first-order partial differential operators:

$$\begin{aligned} &\tilde\partial_1\equiv\partial_t+v_1\partial_x-\bar\alpha',\\ &\tilde\partial_2\equiv\partial_t+v_2\partial_x-\bar\beta',\\ &\tilde\partial_3\equiv\partial_t+v_3\partial_x-\bar\gamma', \end{aligned} \tag{5.10}$$

then (2.9) takes the form

$$\tilde{\partial}_1 D = -a_3 g_2 + a_2 g_3,$$
$$\tilde{\partial}_2 D = -a_1 g_3 + a_3 g_1, \tag{5.11}$$
$$\tilde{\partial}_3 D = -a_2 g_1 + a_1 g_2.$$

On the other hand, it follows from adding together the three equations of (2.10) and using (2.11) that

$$\tilde{\partial}_1 g_1 + \tilde{\partial}_2 g_2 + \tilde{\partial}_3 g_3 = 0. \tag{5.12}$$

However, since $[\tilde{\partial}_1, \tilde{\partial}_2] = 0$, the first two equations of (5.11) imply that

$$\tilde{\partial}_1(-a_1 g_3 + a_3 g_1) = \tilde{\partial}_2(-a_3 g_2 + a_2 g_3)$$

or

$$a_3(\tilde{\partial}_1 g_1 + \tilde{\partial}_2 g_2 + \tilde{\partial}_3 g_3) = (a_1 \tilde{\partial}_1 + a_2 \tilde{\partial}_2 + a_3 \tilde{\partial}_3) g_3. \tag{5.13}$$

The substitution of (5.12) into (5.13) gives

$$(a_1 \tilde{\partial}_1 + a_2 \tilde{\partial}_2 + a_3 \tilde{\partial}_3) g_3 = 0. \tag{5.14}$$

By (5.10) and (5.5), this is precisely

$$(A\partial_t + B\partial_x - C) g_3(x,t) = 0, \tag{5.15}$$

and the most general solution is the same as (5.7), namely,

$$g_3(x,t) = e^{pt+qx} \times \text{function of } \xi, \tag{5.16}$$

with ξ defined by (5.9). By cyclic permutation, $g_1(x,t)$ and $g_2(x,t)$ are also of the same form. [Note that, while p and q are not uniquely determined by (5.7), any shift $(p,q) \to (p',q')$ will merely modify the multiplying function of ξ.] Since (5.7) and (5.16) are of the same form, by (2.6) the three distribution functions are all functions of one variable:

$$f_i(x,t) = \text{function of } (Bt - Ax). \tag{5.17}$$

Solutions of such a simple form do not have enough structure and therefore such models with Eq. (5.1) not trivial are not completely solvable. Appendices A, B, and C are devoted to a further discussion of this class of partially solvable models.

The conclusion is therefore that the parameters of any completely solvable model (2.1) must satisfy (5.2) and (5.3). An alternative way of stating these two conditions is that there is a constant C_1 such that the a's and v's are related by

$$a_1 = C_1(v_2 - v_3),$$
$$a_2 = C_1(v_3 - v_1), \tag{5.18}$$
$$a_3 = C_1(v_1 - v_2).$$

If $C_1 = 0$, then $a_1 = a_2 = a_3 = 0$ and the Boltzmann equation (2.1) is linear. Since linear Boltzmann equations without binary collisions are not difficult to study directly, there is not much point in constructing simplified models in such cases. We therefore exclude this uninteresting case and assume that the constant C_1 is not zero.

Without loss of generality, this constant C_1 can be chosen to be 1. This is most easily accomplished by making a scale transformation on the spatial variable x. By (4.3), this changes v without altering a, α, or f. Since the c_x there can be positive or negative, C_1 can be transformed into 1. Thus, by (5.18),

$$a_1 = v_2 - v_3,$$
$$a_2 = v_3 - v_1, \tag{5.19}$$
$$a_3 = v_1 - v_2.$$

The next step is to get rid of the constants $\bar{\alpha}'$, $\bar{\beta}'$, and $\bar{\gamma}'$ introduced in (2.9). This is accomplished by first observing that, for the same distributions f_1, f_2, and f_3, there are many possible choices of $D(x,t)$, and then choosing an especially simple $D(x,t)$. If (5.2)–(5.4) are considered to be three linear equations for the three a's, then it follows that there are two constants C_2 and C_3 such that

$$\bar{\alpha}' = C_2 + C_3 v_1,$$
$$\bar{\beta}' = C_2 + C_3 v_2, \tag{5.20}$$
$$\bar{\gamma}' = C_2 + C_3 v_3.$$

Therefore the transformation

$$D(x,t) \to D(x,t) e^{C_2 t + C_3 x},$$
$$g_j(x,t) \to g_j(x,t) e^{C_2 t + C_3 x}, \tag{5.21}$$

with $j = 1,2,3$ leaves f_1, f_2, and f_3 unchanged while simplifying the partial differential equations (2.9) and (2.10) for D and g_j to

$$-(\partial_t + v_1\partial_x) D = (v_1 - v_2) g_2 + (v_1 - v_3) g_3,$$
$$-(\partial_t + v_2\partial_x) D = (v_2 - v_3) g_3 + (v_2 - v_1) g_1, \tag{5.22}$$
$$-(\partial_t + v_3\partial_x) D = (v_3 - v_1) g_1 + (v_3 - v_2) g_2$$

and

$$(\partial_t + v_1\partial_x) g_1 = -(\alpha + \alpha') g_1 + \beta g_2 + \gamma' g_3,$$
$$(\partial_t + v_2\partial_x) g_2 = -(\beta + \beta') g_2 + \gamma g_3 + \alpha' g_1, \tag{5.23}$$
$$(\partial_t + v_3\partial_x) g_3 = -(\gamma + \gamma') g_3 + \alpha g_1 + \beta' g_2.$$

The three equations (5.22) are equivalent to the following two:

$$\partial_t D = v_1 g_1 + v_2 g_2 + v_3 g_3 \tag{5.24}$$

and

$$\partial_x D = -(g_1 + g_2 + g_3). \tag{5.25}$$

The six equations of motion (2.9) and (2.10) are therefore reduced to four: (5.23) and (5.24) for the four unknowns D, g_1, g_2, and g_3. Similar to the familiar case of Maxwell's equations, (5.25) is a boundary condition, not an equation of motion. It only remains to show, using (5.23) and (5.24), that, if (5.25) holds at one time, then it holds for all times. This is easily accomplished by adding together the three equations of (5.23):

$$\partial_t(g_1 + g_2 + g_3) + \partial_x(v_1 g_1 + v_2 g_2 + v_3 g_3) = 0. \tag{5.26}$$

The use of (5.24) then gives

$$\partial_t(\partial_x D + g_1 + g_2 + g_3) = 0. \tag{5.27}$$

This proves the assertion.

The most general initial values for the nonlinear Boltzmann equation (2.1) are

$$f_1(x,0)=f_{10}(x),$$
$$f_2(x,0)=f_{20}(x), \tag{5.28}$$
$$f_3(x,0)=f_{30}(x),$$

where $f_{10}(x)$, $f_{20}(x)$, and $f_{30}(x)$ are given. With the f's given by (2.6), this initial-value problem, under the condition (5.18) for the parameters, has been reduced to the linear one of solving the four partial differential equations (5.23) and (5.24) for the four unknowns D, g_1, g_2, and g_3 under the four initial conditions (5.25) and

$$g_1(x,0)=f_{10}(x)D(x,0),$$
$$g_2(x,0)=f_{20}(x)D(x,0), \tag{5.29}$$
$$g_3(x,0)=f_{30}(x)D(x,0).$$

This well-posed problem is to be solved explicitly and exactly in Sec. VII.

VI. EQUILIBRIUM DISTRIBUTIONS

In this section, we initiate the study of equilibrium distributions for the Boltzmann equation (2.1). Although this discussion is entirely elementary, it is nevertheless not easy to present. We choose to follow a comparatively intuitive line of reasoning as follows.

(a) The term "equilibrium distributions" refers to the case where the distributions $f_1(x,t)$, $f_2(x,t)$, and $f_3(x,t)$ are all independent of t, and hence are written as $f_1(x)$, $f_2(x)$, and $f_3(x)$. In this case, (2.1) becomes

$$v_1\frac{df_1}{dx}=f_1[a_3f_2-a_2f_3]-(\alpha+\alpha')f_1+\beta f_2+\gamma' f_3,$$
$$v_2\frac{df_2}{dx}=f_2[a_1f_3-a_3f_1]-(\beta+\beta')f_2+\gamma f_3+\alpha' f_1, \tag{6.1}$$
$$v_3\frac{df_3}{dx}=f_3[a_2f_1-a_1f_2]-(\gamma+\gamma')f_3+\alpha f_1+\beta' f_2.$$

If these three equations are added together, the result is

$$\frac{d}{dx}(v_1f_1+v_2f_2+v_3f_3)=0,$$

or

$$v_1f_1(x)+v_2f_2(x)+v_3f_3(x)=C. \tag{6.2}$$

(b) In this section, we concentrate entirely on the special case where the values of f_1, f_2, and f_3 are furthermore independent of x. They can be interpreted also as the limiting values of $f_1(x)$, $f_2(x)$, and $f_3(x)$ as $x\to\infty$ or $x\to-\infty$.

By (6.1), these f's satisfy simply

$$f_1[a_3f_2-a_2f_3]-(\alpha+\alpha')f_1+\beta f_2+\gamma' f_3=0, \tag{6.3a}$$
$$f_2[a_1f_3-a_3f_1]-(\beta+\beta')f_2+\gamma f_3+\alpha' f_1=0, \tag{6.3b}$$
$$f_3[a_2f_1-a_1f_2]-(\gamma+\gamma')f_3+\alpha f_1+\beta' f_2=0. \tag{6.3c}$$

By (6.2), they also satisfy

$$v_1f_1+v_2f_2+v_3f_3=C. \tag{6.4}$$

It will also be assumed throughout that the relations (5.19) between the three a's and the three v's are valid. It is realized that the problem of equilibrium distributions without these relations is of great interest in its own right. However, such solutions are not relevant to the present completely solvable models of the nonlinear Boltzmann equation.

(c) As a Boltzmann equation, the parameters α, β, γ, α', β', and γ' in (2.1) must satisfy (2.2), i.e., they must be non-negative. In this section, we use a stronger assumption, i.e., they are all positive:

$$\alpha>0,\quad \beta>0,\quad \gamma>0,$$
$$\alpha'>0,\quad \beta'>0,\quad \gamma'>0. \tag{6.5}$$

In the preceding section, it has been assumed that not all three v's are equal. Here we use the stronger assumption that they are all unequal,

$$v_1\neq v_2\neq v_3\neq v_1, \tag{6.6}$$

so that by (5.19)

$$a_1\neq 0,\quad a_2\neq 0,\quad a_3\neq 0. \tag{6.7}$$

Of course the condition (2.4) is not modified

$$f_1\geqslant 0,\quad f_2\geqslant 0,\quad f_3\geqslant 0. \tag{6.8}$$

Actually there is no loss of generality in assuming (6.5) and (6.6). All the other cases can be recovered by taking suitable limits.

(d) By a suitable permutation of the three distributions, we can, without loss of generality, take

$$v_1>v_2>v_3. \tag{6.9}$$

This implies, by (5.19), that

$$a_1>0,\quad a_2<0,\quad a_3>0. \tag{6.10}$$

These signs play a salient role in the following analysis.

(e) Given a_1, a_2, a_3, α, β, γ, α', β', and γ', we define $\mathscr{C}$ to be the set of points (f_1,f_2,f_3) that satisfy (6.3) and (6.8). We proceed to study this $\mathscr{C}$ which is clearly an algebraic curve, or more precisely a part of a real algebraic curve. In particular, we want to determine the number of points there are in the intersection of this $\mathscr{C}$ with the plane (6.4). We first concentrate on obtaining some of the properties of this curve $\mathscr{C}$; only later we introduce the v's and C to define the plane.

Let $\mathscr{R}$ be the set of points (f_1,f_2,f_3) that satisfy (6.8). Then clearly $\mathscr{C}$ is in $\mathscr{R}$. Also the point (0,0,0) is clearly in $\mathscr{C}$.

(f) The first result about $\mathscr{C}$ is that the point (0,0,0) is not an isolated point on $\mathscr{C}$.

This is readily seen by looking at the neighborhood of (0,0,0), i.e., by taking f_1, f_2, and f_3 all small. Thus by (6.3) these small f's satisfy the linear equations

$$-(\alpha+\alpha')f_1+\beta f_2+\gamma' f_3=0,$$

$$\alpha' f_1-(\beta+\beta')f_2+\gamma f_3=0.$$

These two equations imply

$$\frac{f_1}{\beta\gamma+\beta'\gamma'+\beta\gamma'}=\frac{f_2}{\gamma\alpha+\gamma'\alpha'+\gamma\alpha'}=\frac{f_3}{\alpha\beta+\alpha'\beta'+\alpha\beta'}. \tag{6.11}$$

By (6.5) all three denominators are positive. This proves the assertion.

(g) The next elementary property of $\mathscr{C}$ is that, aside from the point (0,0,0), $\mathscr{C}$ does not intersect the boundary of $\mathscr{R}$.

This is readily seen as follows. Let $f_3=0$, then (6.3c) reduces to

$$\alpha f_1+\beta' f_2=0. \tag{6.12}$$

By (6.5), α and β' are positive; by (6.8), f_1 and f_2 are non-negative. Therefore (6.12) can only be satisfied with $f_1=f_2=0$. This proves the assertion.

A picture of this curve $\mathscr{C}$ is beginning to emerge. It has one end point at (0,0,0). Since it cannot intersect the boundary of $\mathscr{R}$ anywhere else, it must approach infinity.

(h) The infinity behavior of $\mathscr{C}$ is

$$f_1\to\infty,$$

$$f_2\to\alpha'/a_3, \tag{6.13}$$

$$f_3\to-\alpha/a_2.$$

That f_1 goes to infinity while f_2 and f_3 do not is closely related to the ordering (6.9) of the velocities. Equivalently, this is due to the signs of the a's, as given by (6.10).

Because of (6.10), the terms quadratic in the f's are of the same sign in both (6.3a) and (6.3c). Therefore it is not possible for more than one of the three f's to approach infinity. Furthermore, in (6.3a), the coefficient of f_2 is $a_3f_1+\beta$ and the coefficient of f_3 is $-a_2f_1+\gamma'$; they are both positive. It is therefore not possible to have either f_2 or f_3 approach infinity with the other two bounded.

Therefore the only possibility is for f_1 to approach infinity. The result (6.13) then follows immediately from (6.3b) and (6.3c).

(i) Let f_2 and f_3 be expressed in terms of f_1. From (6.3b) it follows that

$$f_3=\frac{a_3f_1f_2-\alpha' f_1+(\beta+\beta')f_2}{a_1f_2+\gamma}. \tag{6.14}$$

Substitution into (6.3a) then gives

$$\begin{aligned}&-a_2a_3f_1^2f_2+a_3a_1f_1f_2^2+a_2\alpha' f_1^2-[a_1(\alpha+\alpha')\\&\quad+a_2(\beta+\beta')-a_3(\gamma+\gamma')]f_1f_2+a_1\beta f_2^2\\&\quad-(\gamma\alpha+\gamma'\alpha'+\gamma\alpha')f_1+(\beta\gamma+\beta'\gamma'+\beta\gamma')f_2\\&=0.\end{aligned} \tag{6.15}$$

We shall encounter repeatedly equations of this type. It is thus convenient to introduce the following notation:

Definition:

$$x\Leftrightarrow y \tag{6.16}$$

means that x and y satisfy a third-order polynomial equation where the x^3 and y^3 terms are absent.

The absence of these x^3 and y^3 terms means that, for each x, there are at most two values of y that satisfy the equation, and also that, for each y, there are at most two x's.

An example of this notation is $f_1\Leftrightarrow f_2$.

In (6.15), by (6.8) and (6.10), the coefficient of f_2^2 is

$$a_3a_1f_1+a_1\beta>0,$$

while that of f_2^0 is

$$a_2\alpha' f_1^2-(\gamma\alpha+\gamma'\alpha'+\gamma\alpha')f_1<0$$

for $f_1>0$. Therefore only one of the two possible values of f_2 is positive, and it is given by

$$f_2=\frac{a_2a_3f_1^2+[a_1(\alpha+\alpha')+a_2(\beta+\beta')-a_3(\gamma+\gamma')]f_1-(\beta\gamma+\beta'\gamma'+\beta\gamma')+\sqrt{\Delta}}{2a_1(a_3f_1+\beta)}. \tag{6.17}$$

Substitution into (6.14) then gives

$$f_3=\frac{a_2a_3f_1^2+[-a_1(\alpha+\alpha')+a_2(\beta+\beta')-a_3(\gamma+\gamma')]f_1-(\beta\gamma+\beta'\gamma'+\beta\gamma')+\sqrt{\Delta}}{2a_1(a_2f_1-\gamma')}. \tag{6.18}$$

In (6.17) and (6.18), Δ is given by

$$\begin{aligned}\Delta=&\{a_2a_3f_1^2+[a_1(\alpha+\alpha')+a_2(\beta+\beta')-a_3(\gamma+\gamma')]f_1-(\beta\gamma+\beta'\gamma'+\beta\gamma')\}^2+4a_1f_1(a_3f_1+\beta)\\&\times[(\gamma\alpha+\gamma'\alpha'+\gamma\alpha')-a_2\alpha' f_1].\end{aligned} \tag{6.19}$$

Equations (6.17) and (6.18), with f_1 taking all non-negative values, give an explicit parametrization of $\mathscr{C}$. In particular, it is seen that the algebraic curve $\mathscr{C}$ consists of just one continuous semi-infinite curve that goes from the origin to infinity in the three-dimensional (f_1,f_2,f_3) space.

(j) Asymptotic evaluation of the right-hand expressions of (6.17) and (6.18) for large f_1 gives immediately

$$f_2=\frac{\alpha'}{a_3}-\frac{1}{a_2a_3^2f_1}\left[a_1\alpha\alpha'+a_2\alpha'(\beta+\beta')+a_3\gamma\alpha\right]+O(f_1^{-2}) \tag{6.20}$$

and

$$f_3=-\frac{\alpha}{a_2}-\frac{1}{a_2^2a_3f_1}\left[a_1\alpha\alpha'+a_2\alpha'\beta'+a_3\alpha(\gamma+\gamma')\right]+O(f_1^{-2}). \tag{6.21}$$

In particular, as $f_1\to\infty$, f_2 approaches its limiting value α'/a_3 from below (above) if

$$a_1\alpha\alpha'+a_2\alpha'(\beta+\beta')+a_3\gamma\alpha<0 \quad (>0), \tag{6.22}$$

and f_3 approaches its limiting value $-\alpha/a_2$ from below (above) if

$$a_1\alpha\alpha'+a_2\alpha'\beta'+a_3\alpha(\gamma+\gamma')>0 \quad (<0). \tag{6.23}$$

(k) Considered as functions of f_1, f_2 and/or f_3 is monotonically increasing.

This follows from the fact that, by (6.10),

$$\left[a_1\alpha\alpha'+a_2\alpha'(\beta+\beta')+a_3\gamma\alpha\right]-\left[a_1\alpha\alpha'+a_2\alpha'\beta'+a_3\alpha(\gamma+\gamma')\right]=a_2\alpha'\beta-a_3\gamma'\alpha<0. \tag{6.24}$$

Therefore, by (6.22) and (6.23), it is not possible, as $f_1\to\infty$, for both f_2 and f_3 to approach their respective limiting values from above. At least one of them must approach its limiting value from below.

By (6.15), cyclic permutation, and the definition introduced in subsection (i), we have both $f_1\Leftrightarrow f_2$ and $f_1\Leftrightarrow f_3$. Therefore, if f_2 and/or f_3 approaches its limiting value from below, then f_2 and/or f_3 is a monotonically increasing function of f_1.

(l) So far we have been studying the algebraic curve defined by (6.3), and the additional relation (6.4) between the f's has not been used. In order to analyze the intersection of $\mathscr{C}$ with this plane (6.4), it is convenient to introduce the quantities

$$h_1\equiv -a_3f_2+a_2f_3, \tag{6.25a}$$

$$h_2\equiv -a_1f_3+a_3f_1, \tag{6.25b}$$

$$h_3\equiv -a_2f_1+a_1f_2. \tag{6.25c}$$

They are the combinations that appear in (6.3).

These three h's are not linearly independent:

$$a_1h_1+a_2h_2+a_3h_3=0. \tag{6.26}$$

In (6.15), the third-order terms are

$$-a_2a_3f_1^2f_2+a_3a_1f_1f_2^2=a_3f_1f_2h_3. \tag{6.27}$$

By expressing f_1 in terms of f_2 and h_3, (6.15) can be readily rewritten as a third-order algebraic equation for f_2 and h_3. Furthermore, the terms f_2^3 and h_3^3 are not present. Therefore, by the definition of subsection (i), $f_2\Leftrightarrow h_3$. The same argument also shows that $f_1\Leftrightarrow h_3$. Therefore, by cyclic permutation, the following six relations hold:

$$\begin{aligned}&f_2\Leftrightarrow h_3,\quad f_3\Leftrightarrow h_1,\quad f_1\Leftrightarrow h_2,\\ &f_1\Leftrightarrow h_3,\quad f_2\Leftrightarrow h_1,\quad f_3\Leftrightarrow h_2.\end{aligned} \tag{6.28}$$

(m) Considered as functions of f_1, $-h_1$ and h_3 are both monotonically increasing.

For h_3, this is obvious. Since by (6.10) a_1 is positive while a_2 is negative, h_3 is an increasing function of f_1 for small f_1. By (6.13), it is also an increasing function of f_1 for large f_1. Thus $f_1\Leftrightarrow h_3$ of (6.28) implies that h_3 is a monotonically increasing function of f_1.

The argument for $-h_1$ is more complicated. First, by (6.20), (6.21), and (6.25a), the asymptotic behavior of $-h_1$ for large values of f_1 is given by

$$-h_1=(\alpha+\alpha')-(1/a_2a_3f_1)(a_2\alpha'\beta-a_3\gamma'\alpha)+O(f_1^{-2}). \tag{6.29}$$

This implies that, as $f_1\to\infty$, $-h_1$ always approaches its limiting value $\alpha+\alpha'$ from below.

Secondly, by (k), either f_2 or f_3 or both is a monotonically increasing function of f_1. Suppose f_2 is monotonically increasing. (An entirely similar argument applies in the case of f_3.) In this case, $-h_1$ is an increasing function of f_2 for f_2 near its largest value α'/a_3. Therefore $f_2\Leftrightarrow h_1$ of (6.28) implies that $-h_1$ is a monotonically increasing function of f_2, and hence of f_1. This proves the assertion.

(n) Because of result (m), it is natural to use the variables $-h_1$ and h_3. In terms of these variables [cf. (5.19)]

$$v_1f_1+v_2f_2+v_3f_3=h_3-(v_3/a_2)(h_3-h_1). \tag{6.30}$$

The plane (6.4) is thus

$$v_1h_3-v_3h_1=(v_1-v_3)C. \tag{6.31}$$

Thus the algebraic curve $\mathscr{C}$ is to be projected into this $(-h_1,h_3)$ plane and its intersections with the line (6.31) are to be studied.

This is very fortunate because algebraic curves in two variables are quite simple. In connection with the generalizations discussed in Sec. III, the projection into a two-dimensional plane can also be carried out when the rank of the matrix $\mathscr{A}^{(N)}$ is two, for any finite integer N.

(o) If (6.3) are considered to be linear equations for the f's with the h's of (6.25) as coefficients, then

$$\begin{vmatrix} -h_1-\alpha-\alpha' & \beta & \gamma' \\ \alpha' & -h_2-\beta-\beta' & \gamma \\ \alpha & \beta' & -h_3-\gamma-\gamma' \end{vmatrix}=0. \tag{6.32}$$

When h_2 is expressed in terms of h_1 and h_3 using (6.26), this is the third-order polynomial relating h_1 and h_3, and hence gives the projection of $\mathscr{C}$ into this $(-h_1,h_3)$ plane. In particular,

$$-h_1 \Leftrightarrow h_3. \tag{6.33}$$

(p) Let $\mathscr{C}'$ be the algebraic curve in the $(-h_1,h_3)$ plane given by (6.32) and (6.26). Since the third-order terms are simply $-h_1h_2h_3$, this curve $\mathscr{C}'$ has six asymptotes given by

$$h_1 = -(\alpha+\alpha'), \tag{6.34a}$$

$$h_2 = -(\beta+\beta'), \tag{6.34b}$$

$$h_3 = -(\gamma+\gamma'). \tag{6.34c}$$

For general values of the parameters, $\mathscr{C}'$ connects these six asymptotics pairwise, and hence has three disjoint pieces. Being a third-order polynomial curve, $\mathscr{C}'$ can intersect any straight line at most three times, and furthermore each of the three disjoint pieces can intersect the straight line at most twice.

(q) From (6.11), the slope of $\mathscr{C}'$ at the origin is

$$\left.\frac{d(-h_1)}{dh_3}\right|_0 = \frac{-(v_3-v_1)(\alpha\beta+\alpha'\beta'+\alpha\beta')+(v_1-v_2)(\gamma\alpha+\gamma'\alpha'+\gamma\alpha')}{-(v_3-v_1)(\beta\gamma+\beta'\gamma'+\beta\gamma')+(v_2-v_3)(\gamma\alpha+\gamma'\alpha'+\gamma\alpha')}. \tag{6.35}$$

This is to be compared with the slope of the straight line (6.31). The difference is

$$-\frac{v_1}{v_3}-\left.\frac{d(-h_1)}{dh_3}\right|_0 = -\frac{v_1-v_3}{v_3}\frac{v_1(\beta\gamma+\beta'\gamma'+\beta\gamma')+v_2(\gamma\alpha+\gamma'\alpha'+\gamma\alpha')+v_3(\alpha\beta+\alpha'\beta'+\alpha\beta')}{(v_1-v_3)(\beta\gamma+\beta'\gamma'+\beta\gamma')+(v_2-v_3)(\gamma\alpha+\gamma'\alpha'+\gamma\alpha')}. \tag{6.36}$$

The sign of this quantity depends on the sign of v_3 and that of

$$Z = v_1(\beta\gamma+\beta'\gamma'+\beta\gamma')+v_2(\gamma\alpha+\gamma'\alpha'+\gamma\alpha') + v_3(\alpha\beta+\alpha'\beta'+\alpha\beta'). \tag{6.37}$$

(r) With the conventions on the v's and a's as given by (6.9) and (5.19) (i.e., $C_1 = 1$), the number of solutions (f_1,f_2,f_3) for (6.3) and (6.4) with fixed parameters including C is as follows.

First, if $v_1 = 0$, then (6.4) reduces to $h_1 = C$. Therefore, there is one solution if

$$-(\alpha+\alpha') < C \leqslant 0, \tag{6.38}$$

and no solution otherwise.

Let $v_1 \neq 0$. If the h_3 intercept of the line (6.31) is positive, i.e.,

$$v_1 C > 0, \tag{6.39}$$

then there is one solution.

It remains to consider the case when this intercept is zero or negative. If the three v's have the same sign (including the case $v_3 = 0$), then there is one solution when $C=0$, and no solution if $v_1C < 0$. If the three v's do not have the same sign, then $v_1 > 0$ and $v_3 < 0$, and there are two subcases. If $Z \geqslant 0$, then there is one solution for $C=0$ and no solution for $C<0$. If $Z<0$, there are two solutions for $C=0$, and there are either zero or two solutions for $C<0$. Whether there are zero or two solutions in the last case depends on the sign of the discriminant of a cubic equation.

Under no circumstance are there three solutions to (6.3), (6.4), and (6.8).

VII. SOLUTION OF THE INITIAL-VALUE PROBLEM

In this section, we solve, under the assumption (5.18), the nonlinear Boltzmann equation (2.1) with the initial conditions (5.28).

It has been shown in Sec. V that this is equivalent to solving the partial differential equations (5.23) and (5.24) with the initial conditions (5.25) and (5.29).

The substitution of (5.29) into (5.25) gives

$$\frac{d}{dx}D(x,0) = -[f_{10}(x)+f_{20}(x)+f_{30}(x)]D(x,0) \tag{7.1}$$

or

$$D(x,0) = \exp\left(-\int dx[f_{10}(x)+f_{20}(x)+f_{30}(x)]\right). \tag{7.2}$$

The choice of the constant of integration in the exponent is of no consequence. Therefore, the initial conditions for the linear partial differential equations (5.23) are, for $j=1,2,3$,

$$g_j(x,0) = g_{j0}(x), \tag{7.3}$$

where

$$g_{j0}(x) = f_{j0}(x)\exp\left(-\int dx[f_{10}(x)+f_{20}(x)+f_{30}(x)]\right) \tag{7.4}$$

are known.

Because of the presence of the integral in the exponent, these $g_{j0}(x)$ are expected to be exponentially increasing for large $|x|$ in many cases of interest. The use of Green's functions is thus called for.

Let g and g_0 be the column matrices

$$g(x,t)=\begin{bmatrix} g_1(x,t) \\ g_2(x,t) \\ g_3(x,t) \end{bmatrix} \tag{7.5}$$

and

$$g_0(x,t)=\begin{bmatrix} g_{10}(x,t) \\ g_{20}(x,t) \\ g_{30}(x,t) \end{bmatrix}. \tag{7.6}$$

Define $\mathscr{M}$ to be the 3×3 matrix differential operator with constant coefficients

$$\mathscr{M}=\begin{bmatrix} \partial_t+v_1\partial_x+\alpha+\alpha' & -\beta & -\gamma' \\ -\alpha' & \partial_t+v_2\partial_x+\beta+\beta' & -\gamma \\ -\alpha & -\beta' & \partial_t+v_3\partial_x+\gamma+\gamma' \end{bmatrix}, \tag{7.7}$$

then the problem to be solved is

$$\mathscr{M}g=0, \tag{7.8}$$

with the initial conditions

$$g(x,0)=g_0(x). \tag{7.9}$$

Let I be the 3×3 identity matrix, then the retarded matrix Green's function

$$G=G(x,t)=\begin{bmatrix} G_{11}(x,t) & G_{12}(x,t) & G_{13}(x,t) \\ G_{21}(x,t) & G_{22}(x,t) & G_{23}(x,t) \\ G_{31}(x,t) & G_{32}(x,t) & G_{33}(x,t) \end{bmatrix} \tag{7.10}$$

satisfies the linear partial differential equations

$$\mathscr{M}G=\delta(t)\delta(x)I, \tag{7.11}$$

together with the initial condition

$$G(x,t)=0, \quad \text{for all } t<0. \tag{7.12}$$

In terms of this Green's function, the complete solution to the initial-value problem for the nonlinear Boltzmann equation (2.1) is, for $j=1,2,3$,

$$f_j(x,t)=g_j(x,t)/D(x,t), \tag{7.13}$$

$$g_j(x,t)=\int_{-\infty}^{\infty} dx' \sum_{k=1}^{3} G_{jk}(x-x',t)g_{k0}(x'), \tag{7.14}$$

and

$$D(x,t)=\exp\left(-\int dx \sum_{j=1}^{3} f_{j0}(x)\right) + \int_0^t dt' \sum_{j=1}^{3} v_j g_j(x,t'). \tag{7.15}$$

It remains to find explicitly the Green's function G.

Similar to (5.10), define, for $j=1,2,3$,

$$\partial_j=\mathscr{M}_{jj}, \tag{7.16}$$

or, more explicitly,

$$\begin{aligned} \partial_1&\equiv\partial_t+v_1\partial_x+\alpha+\alpha', \\ \partial_2&\equiv\partial_t+v_2\partial_x+\beta+\beta', \\ \partial_3&\equiv\partial_t+v_3\partial_x+\gamma+\gamma'. \end{aligned} \tag{7.17}$$

From (7.7), define another 3×3 matrix differential operator with constant coefficients

$$\mathscr{M}'=\begin{bmatrix} \partial_2\partial_3-\beta'\gamma & \beta\partial_3+\beta'\gamma' & \gamma'\partial_2+\beta\gamma \\ \alpha'\partial_3+\gamma\alpha & \partial_3\partial_1-\gamma'\alpha & \gamma\partial_1+\gamma'\alpha' \\ \alpha\partial_2+\alpha'\beta' & \beta'\partial_1+\alpha\beta & \partial_1\partial_2-\alpha'\beta \end{bmatrix}, \tag{7.18}$$

then

$$\mathscr{M}\mathscr{M}'=\mathscr{M}'\mathscr{M}=I\mathscr{D}, \tag{7.19}$$

where $\mathscr{D}$ is the third-order differential operator defined by

$$\mathscr{D}=\begin{vmatrix} \partial_t+v_1\partial_x+\alpha+\alpha' & -\beta & -\gamma' \\ -\alpha' & \partial_t+v_2\partial_x+\beta+\beta' & -\gamma \\ -\alpha & -\beta' & \partial_t+v_3\partial_x+\gamma+\gamma' \end{vmatrix}. \tag{7.20}$$

Define a scalar Green's function $G_0=G_0(x,t)$ by

$$\mathscr{D}G_0=\delta(x)\delta(t), \tag{7.21}$$

and the initial condition

$$G_0(x,t)=0, \quad \text{for all } t\leqslant 0. \tag{7.22}$$

Then a comparison of (7.11), (7.21), and (7.19) shows that

$$G=\mathscr{M}'IG_0. \tag{7.23}$$

It remains to find the scalar Green's function G_0. This is

most conveniently accomplished by Fourier transform.

Let

$$G(x,t)=\frac{1}{2\pi}\int_{-\infty}^{\infty} d\xi\, e^{i\xi x}\bar{G}(\xi,t),$$

$$G_0(x,t)=\frac{1}{2\pi}\int_{-\infty}^{\infty} d\xi\, e^{i\xi x}\bar{G}_0(\xi,t), \tag{7.24}$$

$$\bar{\mathscr{M}}=\begin{bmatrix} \partial_t+iv_1\xi+\alpha+\alpha' & -\beta & -\gamma' \\ -\alpha' & \partial_t+iv_2\xi+\beta+\beta' & -\gamma \\ -\alpha & -\beta' & \partial_t+iv_3\xi+\gamma+\gamma' \end{bmatrix}, \tag{7.25}$$

$$\bar{\partial}_j=\bar{\mathscr{M}}_{jj}, \tag{7.26}$$

$$\bar{\mathscr{M}}'=\begin{bmatrix} \bar{\partial}_2\bar{\partial}_3-\beta'\gamma & \beta\bar{\partial}_3+\beta'\gamma' & \gamma'\bar{\partial}_2+\beta\gamma \\ \alpha'\bar{\partial}_3+\gamma\alpha & \bar{\partial}_3\bar{\partial}_1-\gamma'\alpha & \gamma\bar{\partial}_1+\gamma'\alpha' \\ \alpha\bar{\partial}_2+\alpha'\beta' & \beta'\bar{\partial}_1+\alpha\beta & \bar{\partial}_1\bar{\partial}_2-\alpha'\beta \end{bmatrix}, \tag{7.27}$$

and

$$\bar{\mathscr{D}}=\begin{vmatrix} \partial_t+iv_1\xi+\alpha+\alpha' & -\beta & -\gamma' \\ -\alpha' & \partial_t+iv_2\xi+\beta+\beta' & -\gamma \\ -\alpha & -\beta' & \partial_t+iv_3\xi+\gamma+\gamma' \end{vmatrix}, \tag{7.28}$$

then

$$\bar{\mathscr{M}}\bar{\mathscr{M}}'=\bar{\mathscr{M}}'\bar{\mathscr{M}}=I\bar{\mathscr{D}}, \tag{7.29}$$

$$\bar{\mathscr{M}}\bar{G}=\delta(t)I, \tag{7.30}$$

$$\bar{\mathscr{D}}\bar{G}_0=\delta(t), \tag{7.31}$$

$$\bar{G}=\bar{\mathscr{M}}'I\bar{G}_0, \tag{7.32}$$

and

$$\bar{G}(\xi,t)=0, \quad \bar{G}_0(\xi,t)=0, \quad \text{for all } t\leqslant 0. \tag{7.33}$$

Equation (7.31) is merely a third-order ordinary differential equation with constant coefficients, and its solution is, for $t>0$,

$$\bar{G}_0(\xi,t)=A_1(\xi)e^{\omega_1(\xi)t}+A_2(\xi)e^{\omega_2(\xi)t}+A_3(\xi)e^{\omega_3(\xi)t}, \tag{7.34}$$

where $\omega_1(\xi)$, $\omega_2(\xi)$, and $\omega_3(\xi)$ are the roots of the cubic equation

$$\begin{vmatrix} \omega(\xi)+iv_1\xi+\alpha+\alpha' & -\beta & -\gamma' \\ -\alpha' & \omega(\xi)+iv_2\xi+\beta+\beta' & -\gamma \\ -\alpha & -\beta' & \omega(\xi)+iv_3\xi+\gamma+\gamma' \end{vmatrix}=0, \tag{7.35}$$

and the coefficients $A_1(\xi)$, $A_2(\xi)$, and $A_3(\xi)$ are given by

$$\begin{bmatrix} A_1(\xi) \\ A_2(\xi) \\ A_3(\xi) \end{bmatrix}=-\frac{1}{[\omega_1(\xi)-\omega_2(\xi)][\omega_2(\xi)-\omega_3(\xi)][\omega_3(\xi)-\omega_1(\xi)]}\begin{bmatrix} \omega_2(\xi)-\omega_3(\xi) \\ \omega_3(\xi)-\omega_1(\xi) \\ \omega_1(\xi)-\omega_2(\xi) \end{bmatrix}, \tag{7.36}$$

as follows from the boundary condition at $t=0$. Equations (7.24), (7.32), (7.25)–(7.27), and (7.34)–(7.36) give the matrix Green's function $G_{jk}(x,t)$. Together with (7.13)–(7.15), the initial-value problem of the Boltzmann equation (2.1) is solved explicitly and completely.

We add a few remarks about the Green's function $G(x,t)$.

It follows from (7.34)–(7.36) that $\overline{G}_0(\zeta,t)$ [and hence $G(\zeta,t)$] is an analytic function of ζ for any fixed value of t. Furthermore, for large ζ,

$$\begin{aligned}\omega_1(\zeta)&\sim -iv_1\zeta-\alpha-\alpha',\\ \omega_2(\zeta)&\sim -iv_2\zeta-\beta-\beta',\\ \omega_3(\zeta)&\sim -iv_3\zeta-\gamma-\gamma',\end{aligned} \tag{7.37}$$

$$\begin{bmatrix}A_1(\zeta)\\ A_2(\zeta)\\ A_3(\zeta)\end{bmatrix}\sim\frac{1}{\zeta^2(v_1-v_2)(v_2-v_3)(v_3-v_1)}\begin{bmatrix}v_2-v_3\\ v_3-v_1\\ v_1-v_2\end{bmatrix}, \tag{7.38}$$

and hence $\overline{G}_0(\zeta,t)$ is given asymptotically by

$$\begin{aligned}\overline{G}_0(\zeta,t)\sim&[\zeta^2(v_1-v_2)(v_2-v_3)(v_3-v_1)]^{-1}\\ &\times[(v_2-v_3)e^{-(iv_1\zeta+\alpha+\alpha')t}\\ &+(v_3-v_1)e^{-(iv_2\zeta+\beta+\beta')t}\\ &+(v_1-v_2)e^{-(iv_3\zeta+\gamma+\gamma')t}].\end{aligned} \tag{7.39}$$

The substitution of (7.39) into (7.32) gives the asymptotic behavior of $\overline{G}_{jk}(\zeta,t)$ as

$$\begin{aligned}\overline{G}_{11}(\zeta,t)&\sim e^{-(iv_1\zeta+\alpha+\alpha')t},\\ \overline{G}_{22}(\zeta,t)&\sim e^{-(iv_2\zeta+\beta+\beta')t},\\ \overline{G}_{33}(\zeta,t)&\sim e^{-(iv_3\zeta+\gamma+\gamma')t},\end{aligned} \tag{7.40}$$

and

$$\overline{G}_{jk}(\zeta,t)=O(\zeta^{-1}),\quad \text{for } j\neq k.$$

Both (7.39) and (7.40) hold asymptotically for $\zeta\to\infty$ and with any fixed $t>0$.

Because of (7.39) and the analytic property of $\overline{G}_0(\zeta,t)$, the contour of integration in (7.24) can be closed at infinity when $x-v_1t$, $x-v_2t$, and $x-v_3t$ all have the same sign. With the convention (6.9), $v_1>v_2>v_3$, we get, for $t>0$,

$$G_0(x,t)=0,\ \text{and}\ G(x,t)=0, \tag{7.41}$$

when

$$x-v_1t>0, \tag{7.42}$$

and when

$$x-v_3t<0. \tag{7.43}$$

This is an expression of the fact that the Boltzmann equation (2.1) has limited velocities for the particles colliding with each other and the background gas. Substitution into (7.14) then gives, for $j=1,2,3$,

$$g_j(x,t)=\int_{x-v_1t-}^{x-v_3t+}dx'\sum_{k=1}^{3}G_{jk}(x-x',t)g_{k0}(x'). \tag{7.44}$$

Here the $-$ and $+$ signs in the range of integration merely express the fact that there are δ functions in G_{jk} at the end points of the range of integration that must be included.

For purposes of numerical integration, it is desirable to write out explicitly these δ functions in $G_{jk}(x,t)$. They are readily available from (7.40). Define

$$H=H(x,t)=\begin{bmatrix}H_{11}(x,t) & H_{12}(x,t) & H_{13}(x,t)\\ H_{21}(x,t) & H_{22}(x,t) & H_{23}(x,t)\\ H_{31}(x,t) & H_{32}(x,t) & H_{33}(x,t)\end{bmatrix}, \tag{7.45}$$

so that

$$G(x,t)=H(x,t)+\begin{bmatrix}e^{-(\alpha+\alpha')t}\delta(x-v_1t) & 0 & 0\\ 0 & e^{-(\beta+\beta')t}\delta(x-v_2t) & 0\\ 0 & 0 & e^{-(\gamma+\gamma')t}\delta(x-v_3t)\end{bmatrix}, \tag{7.46}$$

then $H(x,t)$ contains no δ-function term.

We summarize the results of this section. $\overline{G}_0(\zeta,t)$ is given by (7.34)–(7.36), and then $\overline{G}(\zeta,t)$ is given by (7.32) and (7.27). Let

$$\overline{H}(\zeta,t)=\overline{G}(\zeta,t)-\begin{bmatrix}e^{-(iv_1\zeta+\alpha+\alpha')t} & 0 & 0\\ 0 & e^{-(iv_2\zeta+\beta+\beta')t} & 0\\ 0 & 0 & e^{-(iv_3\zeta+\gamma+\gamma')t}\end{bmatrix}, \tag{7.47}$$

and

$$H(x,t)=\frac{1}{2\pi}\int_{-\infty}^{\infty} d\zeta\, e^{i\zeta x}\overline{H}(\zeta,t), \tag{7.48}$$

then

$$g_j(x,t)=e^{-\nu_j t}g_{j0}(x-v_j t) + \int_{x-v_1 t}^{x-v_3 t} dx' \sum_{k=1}^{3} H_{jk}(x-x',t)g_{k0}(x'), \tag{7.49}$$

where g_{j0} is given by (7.4) and

$$\begin{aligned}\nu_1&=\alpha+\alpha',\\ \nu_2&=\beta+\beta',\\ \nu_3&=\gamma+\gamma'.\end{aligned} \tag{7.50}$$

(These are Greek ν's in the exponential in (7.49), not to be confused with the velocities v_j)

Finally, in terms of this $g_j(x,t)$, the time evolution of the distributions $f_j(x,t)$ is given by (7.13) and (7.15).

VIII. SIMPLE ILLUSTRATIONS

A program has been initiated to study numerically the present exact solutions of the Boltzmann equation. The results will be reported on in a separate publication.[11] However, in order to develop some intuition for these models, we include also here a couple of examples that illustrate the most simple properties.

In Figs. 1, 2, and 3 we consider three examples (example A, example B, and example C) given by the following initial distributions:

example A:

$$\begin{aligned}f_{10}(x)&=e^{-(x+1)^2},\\ f_{20}(x)&=e^{-(x-1)^2},\\ f_{30}(x)&=0;\end{aligned} \tag{8.1}$$

example B:

$$\begin{aligned}f_{10}(x)&=1/(1+e^x),\\ f_{20}(x)&=e^x/(1+e^x),\\ f_{30}(x)&=0;\end{aligned} \tag{8.2}$$

example C:

$$\begin{aligned}f_{10}(x)&=1-e^{-x^2},\\ f_{20}(x)&=e^{-x^2},\\ f_{30}(x)&=0.\end{aligned} \tag{8.3}$$

Thus, for these three examples, there are initially no slow particles, $f_3(x,t)\equiv 0$. The distributions are shown in Figs. 1(a), 2(a), and 3(a).

We consider the velocities

$$v_1=4,\quad v_2=2,\quad v_3=1. \tag{8.4}$$

The time evolution will be indicated for three choices of the parameters α, β, γ, α', β', and γ'.

With

$$\alpha=\beta=\gamma=\alpha'=\beta'=\gamma'=0, \tag{8.5}$$

the time development is given entirely by the nonlinear terms. Figures 1(b), 2(b), and 3(b) show how particles get transferred from a distribution of slow ones, to another of faster ones, at rates given by the overlap of the two distributions. If we define

$$N_j(t)\equiv\int_{-\infty}^{\infty} dx\, f_j(x,t), \tag{8.6}$$

we see that $N_1(t)$ increases, while $N_2(t)$ decreases with increasing t, with $N_3(t)$ remaining zero.

With the choice (8.5), the $H(x,t)$ of Eqs. (7.45)–(7.49) vanishes. Thus, only the δ-function part of the Green's function contributes, and the numerical evaluation of the $g_j(x,t)$ of Eq. (7.49) becomes trivial. [However, the numerical integration (7.15) that gives $D(x,t)$ still has to be performed.]

When the coefficients of the linear terms are nonzero, particles also get transferred from one distribution to another, irrespectively of the overlap. (The linear collision terms may be interpreted as representing interactions with a background gas.) In Figs. 1(c), 2(c), and 3(c) we show the initial time evolution ($t=0$, 0.025, 0.050,...,0.200) for

$$\begin{aligned}\alpha&=\alpha'=1,\\ \beta&=\beta'=\gamma=\gamma'=0.\end{aligned} \tag{8.7}$$

With this choice, the only linear collision terms that are active, are those by which the distribution $f_1(x,t)$ loses particles to $f_2(x,t)$ and $f_3(x,t)$. [Compare Eq. (2.1).] With these parameters, $f_3(x,t)$ thus becomes nonzero. We see from the figures that it only "inherits" particles from $f_1(x,t)$, not from $f_2(x,t)$.

In this case, the cubic equation (7.35), which governs the time evolution of the Green's function, becomes trivial and the $\omega_i(\zeta)$ become linear expressions in ζ. As a result, the Green's function can be determined analytically.

The last set of parameters considered here, is

$$\alpha=\beta=\gamma=\alpha'=\beta'=\gamma'=1, \tag{8.8}$$

i.e., all linear terms are present, with equal strength. The initial time evolutions ($t=0,0.025,0.050,...,0.200$) are shown in Figs. 1(d), 2(d), and 3(d). A rich structure is seen to emerge.

IX. SUMMARY

The models described in the present paper have the property that particles are transferred between three distributions by two independent and sometimes competing mechanisms.

(1) Particles are transferred from a distribution of slow ones, to another of faster ones, at rates given by the difference in velocity, and the overlap of the two distributions, according to the nonlinear collision terms in Eq. (2.1) [see also (5.18)].

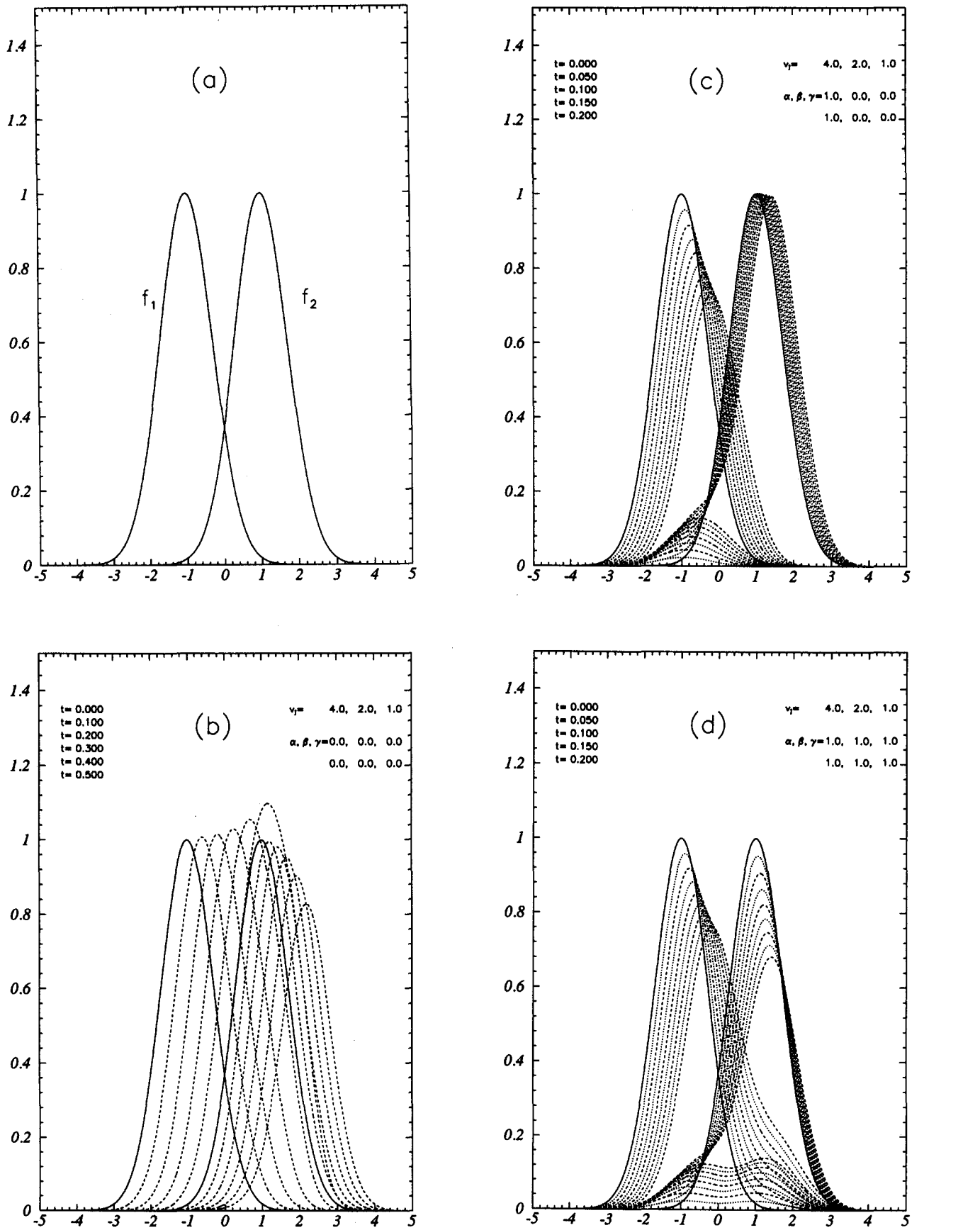

FIG. 1. Time evolutions for initial conditions specified by example A of Eq. (8.1), and $v_1 = 4$, $v_2 = 2$, and $v_3 = 1$. (a) Initial ($t=0$) distributions. (b) Only the *nonlinear* collision terms are present. The distributions are shown for $t=0$ [solid, same as (a)], $t=0.1$, 0.2, 0.3, 0.4, and 0.5 (dashed, increasing times to the right). Note that f_3 remains zero. (c) Two linear collision terms are present, $\alpha = \alpha' = 1$, whereas $\beta = \beta' = \gamma = \gamma' = 0$, for $t=0$ (solid), $t=0.025$, 0.005, 0.075,... up to 0.20 (alternately dotted and dashed, increasing times to the right). Note that f_3 is fed by f_1. (d) Same as (c) except that all linear collision terms are present with equal strength, $\alpha = \alpha' = \beta = \beta' = \gamma = \gamma' = 1$. Now f_3 is fed by both f_1 and f_2.

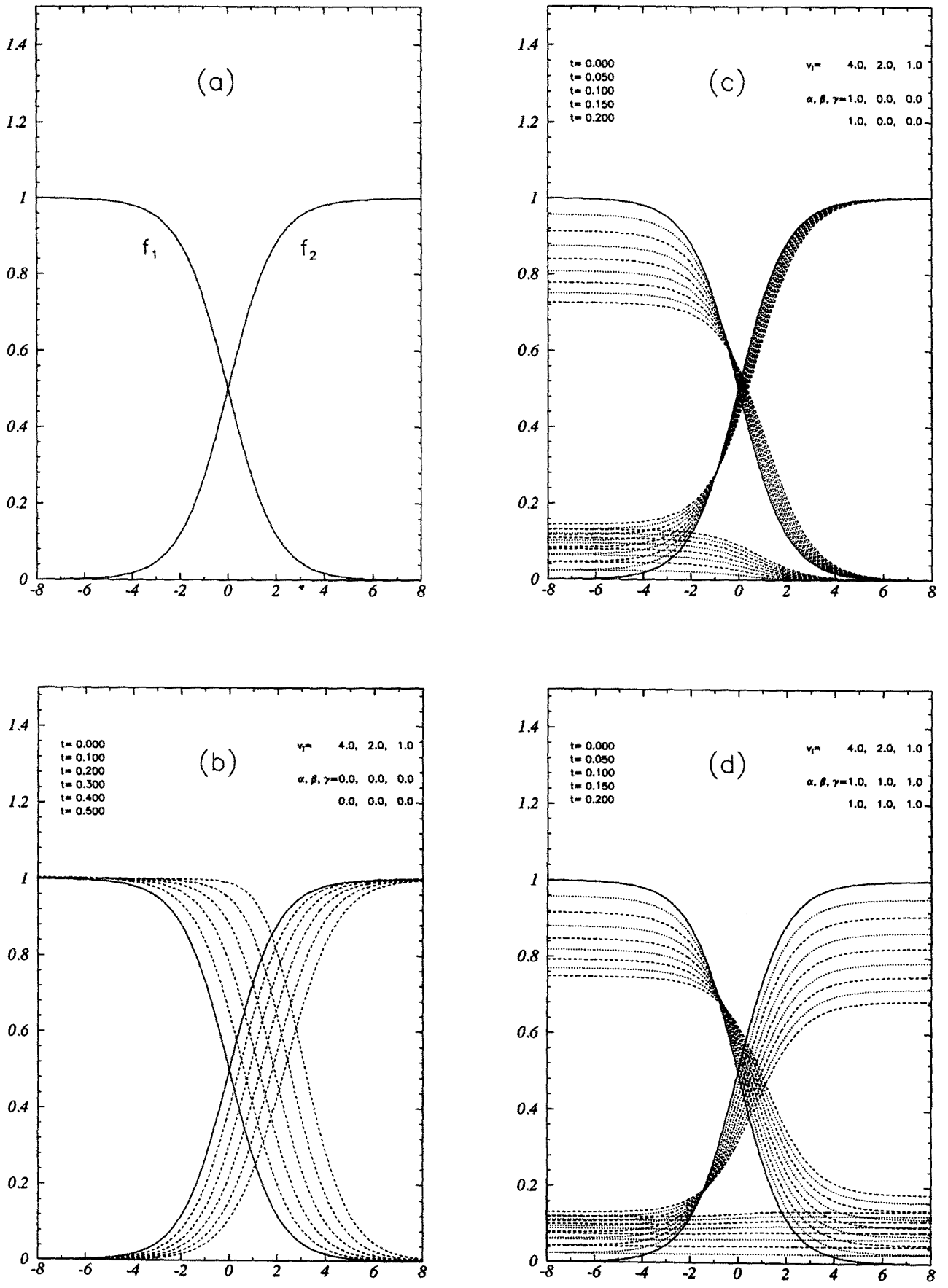

FIG. 2. Similar to Fig. 1, for initial conditions given by example B of Eq. (8.2).

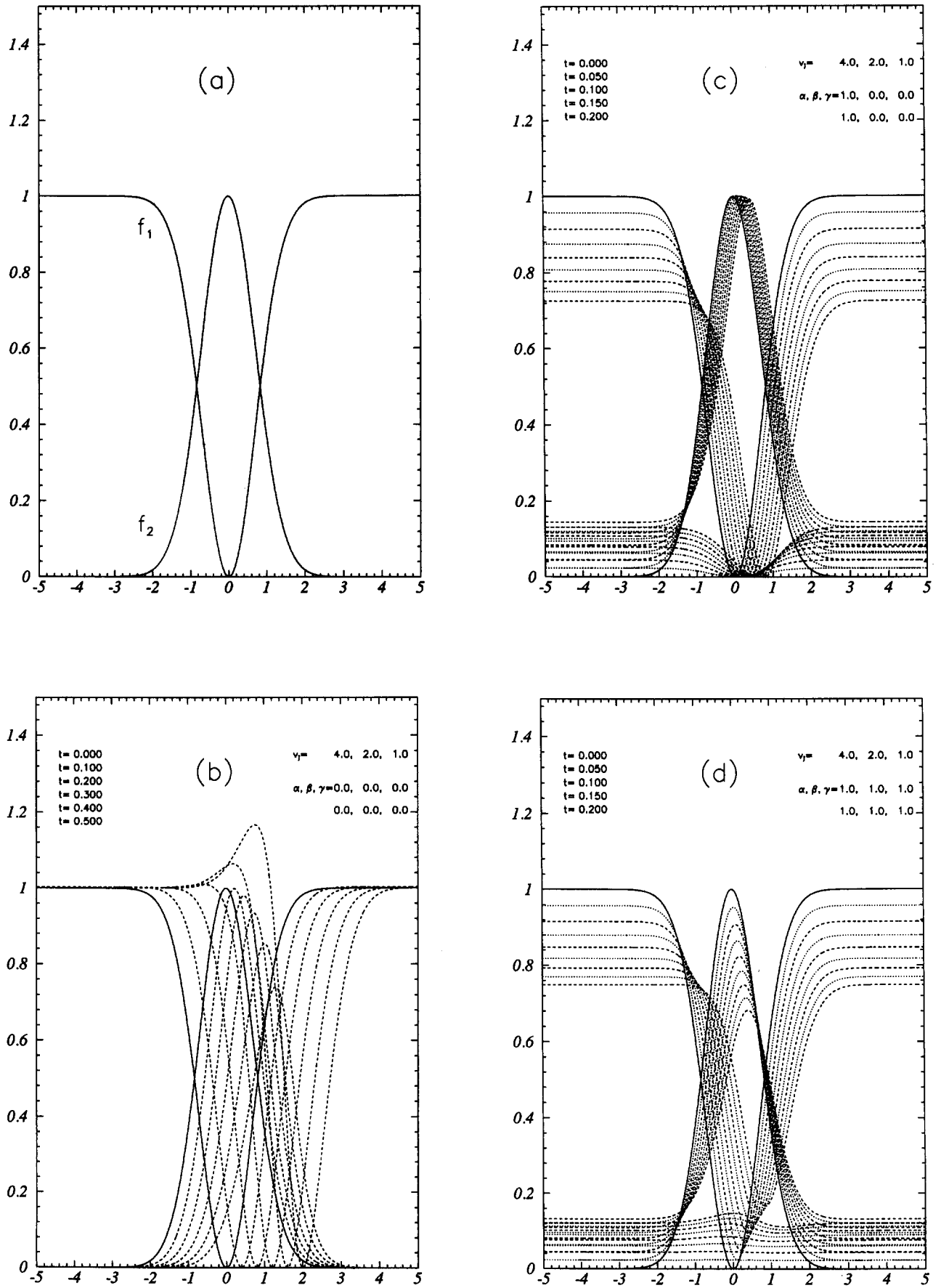

FIG. 3. Similar to Fig. 1, for initial conditions given by example C of Eq. (8.3).

(2) Particles are transferred between the three distributions at rates given by the linear terms, i.e., irrespectively of any overlap.

Various different time evolutions will thus occur, depending upon how the two mechanisms interfere, which is determined by the initial distributions and by the parameters of the Boltzmann equation. Some of these will be further explored in a separate paper.[11]

ACKNOWLEDGMENTS

We thank Professor Harry Lehmann and Professor Roberto Peccei for very helpful discussions. A substantial part of this work was carried out in 1985 when one of us (TTW) was at Deutsches Elektronen-Synchrotron DESY, Hamburg, Germany. He is grateful to the Alexander von Humboldt Foundation for a Humboldt award. Finally, it is a pleasure to thank P. S. Iversen and A. K. Topphol for patiently explaining how to use the plotting facilities.

The work of PO was supported in part by the Norwegian Research Council for Science and the Humanities. The work of TTW was supported in part by the U.S. Department of Energy under Grant DE-FG02-84ER 40158.

APPENDIX A: MODELS WITH INITIAL DATA GIVEN BY ONE ARBITRARY FUNCTION

There are some classes of models where the three distribution functions are given by functions of *one* argument, $\phi_i(Bt - Ax)$ [see Eq. (5.17)]. The initial values must then be given by the same functions, evaluated at $t=0$. Not including cases where one of the distribution functions vanishes identically, we have *a priori* the following possibilities:

$$\text{Class A1:}\quad f_1(x,t)=\phi_1(Bt-Ax),\quad f_2(x,t)=c_2,\quad f_3(x,t)=c_3; \tag{A1}$$

$$\text{Class A2:}\quad f_1(x,t)=\phi_1(Bt-Ax),\quad f_2(x,t)=\phi_2(Bt-Ax),\quad f_3(x,t)=c_3; \tag{A2}$$

$$\text{Class A3:}\quad f_1(x,t)=\phi_1(Bt-Ax),\quad f_2(x,t)=\phi_2(Bt-Ax),\quad f_3(x,t)=\phi_3(Bt-Ax), \tag{A3}$$

where cyclic permutations of classes A1 and A2 are not distinct classes, and with c_2 and c_3 nonzero constants. We leave open the possibility that some of the ϕ_i might be proportional.

A lengthy analysis is required in order to determine which of these classes are permitted (class A3 does not exist), and whether there are any conditions on the ϕ_i. In this appendix we first give some general results, and subsequently present models where the ϕ_i can be expressed in terms of *one* arbitrary function, while in Appendix B we present models where *two* arbitrary functions are accommodated.

1. Preliminaries

We start out defining three auxiliary functions,

$$F_1=(\partial_t+v_1\partial_x-\bar{\alpha}')D,\quad F_2=(\partial_t+v_2\partial_x-\bar{\beta}')D,\quad F_3=(\partial_t+v_3\partial_x-\bar{\gamma}')D. \tag{A4}$$

Equation (5.6) is then equivalent to the condition

$$a_1F_1+a_2F_2+a_3F_3=0. \tag{A5}$$

In terms of $D_0(\xi)$ [cf. Eqs. (5.7) and (5.9)], these three quantities are

$$F_1=[(p+v_1q-\bar{\alpha}')D_0+(B-v_1A)D_0']e^{pt+qx},\quad F_2=[(p+v_2q-\bar{\beta}')D_0+(B-v_2A)D_0']e^{pt+qx},\quad F_3=[(p+v_3q-\bar{\gamma}')D_0+(B-v_3A)D_0']e^{pt+qx}, \tag{A6}$$

with $D_0'=(d/d\xi)D_0(\xi)$.

Returning now to Eq. (2.9), we find that the g_i must satisfy

$$a_3g_2-a_2g_3=-F_1,\quad a_1g_3-a_3g_1=-F_2,\quad a_2g_1-a_1g_2=-F_3. \tag{A7}$$

Assuming all $a_i\neq 0$ (the case of one or two of the a_i being zero is discussed in Appendix C), the most general solution to Eq. (A7) is given by

$$g_1=\frac{1}{3}\left(\frac{F_2}{a_3}-\frac{F_3}{a_2}\right)+a_1g(x,t)\equiv g_{10}+a_1g(x,t),\quad g_2=\frac{1}{3}\left(\frac{F_3}{a_1}-\frac{F_1}{a_3}\right)+a_2g(x,t)\equiv g_{20}+a_2g(x,t),\quad g_3=\frac{1}{3}\left(\frac{F_1}{a_2}-\frac{F_2}{a_1}\right)+a_3g(x,t)\equiv g_{30}+a_3g(x,t), \tag{A8}$$

as is readily confirmed by substitution.

The function $g(x,t)$ is constrained by the fact that the g_i have to satisfy Eq. (2.10). Substituting Eq. (A8) into (2.10), we find the following three equations:

$$\begin{aligned}(\partial_t+v_1\partial_x)(g_{10}+a_1g)&=-(\alpha+\alpha'-\bar{\alpha}')(g_{10}+a_1g)\\&\quad+\beta(g_{20}+a_2g)\\&\quad+\gamma'(g_{30}+a_3g),\\(\partial_t+v_2\partial_x)(g_{20}+a_2g)&=-(\beta+\beta'-\bar{\beta}')(g_{20}+a_2g)\\&\quad+\gamma(g_{30}+a_3g)\\&\quad+\alpha'(g_{10}+a_1g),\\(\partial_t+v_3\partial_x)(g_{30}+a_3g)&=-(\gamma+\gamma'-\bar{\gamma}')(g_{30}+a_3g)\\&\quad+\alpha(g_{10}+a_1g)\\&\quad+\beta'(g_{20}+a_2g).\end{aligned} \tag{A9}$$

Before studying these equations in detail, we note that their sum yields [cf. Eq. (5.15)]

$$(A\partial_t + B\partial_x - C)g(x,t)=0.$$

This is the same as Eq. (5.6), i.e., $g(x,t)$ and $D(x,t)$ must satisfy the *same* partial differential equation.

In analogy with Eq. (5.7), we introduce a function of *one* variable, $g_0(\xi)$, by the ansatz

$$g(x,t)=e^{pt+qx}g_0(Bt-Ax). \tag{A10}$$

If we now substitute this into Eq. (A9), and substitute for the g_{10}, g_{20}, and g_{30} in terms of $D_0(\xi)$, according to Eqs. (A8) and (A6), we obtain three coupled *ordinary* differential equations for $g_0(\xi)$ and $D_0(\xi)$. These have the forms

$$\begin{aligned}
&\left(P_1\frac{d}{d\xi}+Q_1\right)g_0+\left(R_1\frac{d^2}{d\xi^2}+S_1\frac{d}{d\xi}+T_1\right)D_0=0,\\
&\left(P_2\frac{d}{d\xi}+Q_2\right)g_0+\left(R_2\frac{d^2}{d\xi^2}+S_2\frac{d}{d\xi}+T_2\right)D_0=0,\\
&\left(P_3\frac{d}{d\xi}+Q_3\right)g_0+\left(R_3\frac{d^2}{d\xi^2}+S_3\frac{d}{d\xi}+T_3\right)D_0=0,
\end{aligned} \tag{A11}$$

with

$$P_1=a_1(B-v_1A), \tag{A12a}$$

$$Q_1=a_1(p+v_1q+\bar{\alpha})-a_2\beta-a_3\gamma', \tag{A12b}$$

$$R_1=\frac{1}{3}(B-v_1A)\left(\frac{B-v_2A}{a_3}-\frac{B-v_3A}{a_2}\right), \tag{A12c}$$

$$\begin{aligned}
S_1=\frac{1}{3}\Bigg[&(B-v_1A)\left(\frac{p+v_2q-\bar{\beta}'}{a_3}-\frac{p+v_3q-\bar{\gamma}'}{a_2}\right)\\
&+(p+v_1q+\bar{\alpha})\left(\frac{B-v_2A}{a_3}-\frac{B-v_3A}{a_2}\right)\\
&-\beta\left(\frac{B-v_3A}{a_1}-\frac{B-v_1A}{a_3}\right)\\
&-\gamma'\left(\frac{B-v_1A}{a_2}-\frac{B-v_2A}{a_1}\right)\Bigg],
\end{aligned} \tag{A12d}$$

$$\begin{aligned}
T_1=\frac{1}{3}\Bigg[&(p+v_1q+\bar{\alpha})\left(\frac{p+v_2q-\bar{\beta}'}{a_3}-\frac{p+v_3q-\bar{\gamma}'}{a_2}\right)\\
&-\beta\left(\frac{p+v_3q-\bar{\gamma}'}{a_1}-\frac{p+v_1q-\bar{\alpha}'}{a_3}\right)\\
&-\gamma'\left(\frac{p+v_1q-\bar{\alpha}'}{a_2}-\frac{p+v_2q-\bar{\beta}'}{a_1}\right)\Bigg],
\end{aligned} \tag{A12e}$$

and P_2, Q_2,...,T_3 obtained by cyclic permutations. (The corresponding results for $a_1 = 0$ and for $a_1 = a_2 = 0$ are given in Appendix C.)

Because of (5.8) and (2.11), it follows that

$$\sum_{j=1}^{3}P_j=\sum_{j=1}^{3}Q_j=\sum_{j=1}^{3}R_j=\sum_{j=1}^{3}S_j=\sum_{j=1}^{3}T_j=0. \tag{A13}$$

Therefore, one of the three equations in (A11) is redundant, and it suffices to consider the *two* equations

$$\begin{aligned}
&(P_1\mathscr{D}+Q_1)g_0+(R_1\mathscr{D}^2+S_1\mathscr{D}+T_1)D_0=0,\\
&(P_2\mathscr{D}+Q_2)g_0+(R_2\mathscr{D}^2+S_2\mathscr{D}+T_2)D_0=0,
\end{aligned} \tag{A14}$$

where $\mathscr{D}\equiv(d/d\xi)$. The above differential operators commute, so we can apply $(P_2\mathscr{D}+Q_2)$ to the first equation, and $(P_1\mathscr{D}+Q_1)$ to the second, thereby eliminating g_0. This yields

$$UD_0=0, \tag{A15}$$

with

$$\begin{aligned}
U=&(P_1\mathscr{D}+Q_1)(R_2\mathscr{D}^2+S_2\mathscr{D}+T_2)\\
&-(P_2\mathscr{D}+Q_2)(R_1\mathscr{D}^2+S_1\mathscr{D}+T_1).
\end{aligned}$$

Similarly, one finds

$$Ug_0=0. \tag{A16}$$

Hence, D_0 and g_0 satisfy the *same* ordinary differential equation with constant coefficients. This differential equation has to be trivial. Otherwise, D_0 and g_0 would be sums of exponentials, and more general initial ($t=0$) conditions could not be satisfied. We must therefore require

$$U\equiv 0.$$

More explicitly, this means

$$P_1R_2=P_2R_1, \tag{A17a}$$

$$P_1S_2+Q_1R_2=P_2S_1+Q_2R_1, \tag{A17b}$$

$$P_1T_2+Q_1S_2=P_2T_1+Q_2S_1, \tag{A17c}$$

$$Q_1T_2=Q_2T_1. \tag{A17d}$$

These are the consistency conditions for "model 1." When these conditions are satisfied, $D_0(\xi)$ is *arbitrary*, with $g_0(\xi)$ determined by $D_0(\xi)$ according to (A14).

Alternatively, one could require

$$P_i=Q_i=R_i=S_i=T_i=0, \quad i=1,2.$$

Then $g_0(\xi)$ and $D_0(\xi)$ would *both* be arbitrary. This leads to "model 2," which is discussed in Appendix B.

2. Model 1 with $P_2\neq 0$, $P_3\neq 0$

By Eq. (A13), we must have

$$P_1+P_2+P_3=0.$$

There are two possibilities, either all P_i vanish, or at least two of them are nonzero. We first consider the latter possibility. By a suitable labeling, these nonvanishing ones can be chosen as P_2 and P_3. Hence,

$$P_2\neq 0, \quad P_3\neq 0,$$

and by (A12a) we must have

$$a_2\neq 0, \quad a_3\neq 0, \quad B-v_2A\neq 0, \quad B-v_3A\neq 0. \tag{A18}$$

Irrespectively of whether or not $a_1\neq 0$, the consistency condition (A17a) is then found to lead to the requirement that

$$B-v_1A=0, \tag{A19}$$

i.e.,

$$v_1=(a_2v_2+a_3v_3)/(a_2+a_3). \tag{A20}$$

Equations (A18) and (A19) imply that all velocities have to be different:

$$v_1\neq v_2, \quad v_2\neq v_3, \quad v_3\neq v_1. \tag{A21}$$

Furthermore, Eq. (A19) implies [see Eqs. (A12a) and (A12c)]

$$P_1=0, \tag{A22a}$$

$$R_1=0. \tag{A22b}$$

Equation (A17b) then simplifies to

$$Q_1R_2=P_2S_1,$$

which leads to the following condition involving p and q:

$$p+v_1q+\bar{\alpha}=0. \tag{A23}$$

In order to proceed with the analysis of Eqs. (A17c) and (A17d), we have to consider separately the two cases of $a_1\neq 0$ and $a_1=0$. For both cases, the consistency condition (A17c), which by (A22a) reduces to

$$Q_1S_2=P_2T_1+Q_2S_1,$$

yields

$$a_2\beta\alpha'=a_3\gamma'\alpha=K, \tag{A24}$$

where K is some constant.

For $a_1=0$, the original equations (2.1) are symmetric under the interchange $f_2\leftrightarrow f_3$, provided one also interchanges

$$\begin{aligned} &v_2\leftrightarrow v_3,\\ &a_2\leftrightarrow -a_3,\\ &\beta\leftrightarrow\gamma',\\ &\beta'\leftrightarrow\gamma,\\ &\alpha\leftrightarrow\alpha'. \end{aligned} \tag{A25}$$

Equation (A24) is seen to satisfy this symmetry requirement.

The remaining consistency condition, Eq. (A17d), is more involved. We first consider $a_1\neq 0$. Invoking Eq. (5.8), which can be written out as

$$\begin{aligned} &a_1(p+v_1q-\bar{\alpha}')+a_2(p+v_2q-\bar{\beta}')\\ &\quad+a_3(p+v_3q-\bar{\gamma}')=0, \end{aligned} \tag{A26}$$

and furthermore Eqs. (A19) and (A23), we find

$$\begin{aligned} &(\alpha+\alpha')(\beta+\beta')\gamma'-(p+v_1q-\bar{\alpha}')\beta\gamma\\ &\quad+(p+v_2q-\bar{\beta}')\alpha\gamma'+(p+v_3q-\bar{\gamma}')\alpha'\beta=0. \end{aligned}$$

We here use Eqs. (A24) and (A26) to combine the last two terms, obtaining thus

$$(\alpha+\alpha')[\beta\gamma+\beta\gamma'+\beta'\gamma'+Ka_1/a_2a_3]=0. \tag{A27}$$

A corresponding analysis for the case of $a_1=0$ yields the "same" result, i.e., one may simply substitute $a_1=0$ in Eq. (A27).

There are two solutions to Eq. (A27):

$$\text{(i)}\quad \alpha=\alpha'=0; \quad a_1,a_2,a_3,\beta,\beta',\gamma,\gamma' \quad \text{arbitrary}, \tag{A28}$$

$$\text{(ii)}\quad \beta\gamma+\beta\gamma'+\beta'\gamma'+Ka_1/a_2a_3=0; \quad \alpha,\alpha' \quad \text{subject to (A24)}. \tag{A29}$$

In both cases, v_1, v_2, and v_3 are subject to conditions (A20) and (A21).

If in the latter case $a_1=0$, it suffices to have

$$\beta=\beta'=0, \quad \text{with } a_3\gamma'\alpha=0,$$

or

$$\beta=\gamma'=0, \quad \text{or } \gamma=\gamma'=0, \quad \text{with } a_2\beta\alpha'=0.$$

We next study the corresponding distribution functions f_i.

3. Model 1 with $P_2\neq 0$, $P_3\neq 0$, and $a_1\neq 0$

We first consider

Case I:

$$a_1\neq 0.$$

The g_i are then given by the general expression, Eq. (A8), with the auxiliary functions F_i given by Eqs. (A6) and (A19) as:

$$\begin{aligned} F_1&=(p+v_1q-\bar{\alpha}')D_0e^{pt+qx},\\ F_2&=[(p+v_2q-\bar{\beta}')D_0+(B-v_2A)D_0']e^{pt+qx},\\ F_3&=[(p+v_3q-\bar{\gamma}')D_0+(B-v_3A)D_0']e^{pt+qx}. \end{aligned}$$

It follows from Eq. (2.6) that

$$\begin{aligned} f_1(x,t)&=\frac{1}{3}\left[\frac{p+v_2q-\bar{\beta}'}{a_3}-\frac{p+v_3q-\bar{\gamma}'}{a_2}+\left(\frac{B-v_2A}{a_3}-\frac{B-v_3A}{a_2}\right)\frac{D_0'(\xi)}{D_0(\xi)}\right]+a_1\frac{g_0(\xi)}{D_0(\xi)},\\ f_2(x,t)&=\frac{1}{3}\left[\frac{p+v_3q-\bar{\gamma}'}{a_1}-\frac{p+v_1q-\bar{\alpha}'}{a_3}+\frac{B-v_3A}{a_1}\frac{D_0'(\xi)}{D_0(\xi)}\right]+a_2\frac{g_0(\xi)}{D_0(\xi)},\\ f_3(x,t)&=\frac{1}{3}\left[\frac{p+v_1q-\bar{\alpha}'}{a_2}-\frac{p+v_2q-\bar{\beta}'}{a_1}-\frac{B-v_2A}{a_1}\frac{D_0'(\xi)}{D_0(\xi)}\right]+a_3\frac{g_0(\xi)}{D_0(\xi)}. \end{aligned} \tag{A30}$$

The functions $g_0(\xi)$ and $D_0(\xi)$ are related by Eq. (A14). For the case under consideration [see Eqs. (A22a) and (A22b)]:

$$g_0(\xi)=-\frac{1}{Q_1}\left[S_1\frac{d}{d\xi}+T_1\right]D_0(\xi). \tag{A31}$$

By Eqs. (A12b), (A12d), (A12e), (A19), and (A23),

$$Q_1=-a_2\beta-a_3\gamma',$$

$$S_1=-\frac{\beta}{3a_1}(B-v_3A)+\frac{\gamma'}{3a_1}(B-v_2A),$$

$$T_1=-\frac{1}{3}\left[\beta\left(\frac{p+v_3q-\bar{\gamma}'}{a_1}-\frac{p+v_1q-\bar{\alpha}'}{a_3}\right)+\gamma'\left(\frac{p+v_1q-\bar{\alpha}'}{a_2}-\frac{p+v_2q-\bar{\beta}'}{a_1}\right)\right]. \tag{A32}$$

Quite generally, for model 1 with $P_2\neq 0$, $P_3\neq 0$, and $a_1\neq 0$, we then have

$$\frac{g_0(\xi)}{D_0(\xi)}=\frac{1}{3a_1}\frac{-\beta(B-v_3A)+\gamma'(B-v_2A)}{a_2\beta+a_3\gamma'}\frac{D_0'(\xi)}{D_0(\xi)}-\frac{1}{3(a_2\beta+a_3\gamma')}\times\left[\beta\left(\frac{p+v_3q-\bar{\gamma}'}{a_1}-\frac{p+v_1q-\bar{\alpha}'}{a_3}\right)+\gamma'\left(\frac{p+v_1q-\bar{\alpha}'}{a_2}-\frac{p+v_2q-\bar{\beta}'}{a_1}\right)\right]. \tag{A33}$$

Therefore, the distribution functions depend on *one* "arbitrary" function $\phi(\xi)$ as follows:

$$f_1(x,t)=f_{10}+\eta_1\phi(\xi),$$
$$f_2(x,t)=f_{20}+\eta_2\phi(\xi), \tag{A34}$$
$$f_3(x,t)=f_{30}+\eta_3\phi(\xi).$$

Substituting (A33) into (A30), and using (5.8) and (A23), we find

$$\eta_1=(a_2\beta+a_3\gamma')^{-1}[-\beta(B-v_3A)+\gamma'(B-v_2A)],$$
$$\eta_2=0,$$
$$\eta_3=0.$$

4. Model 1 with $P_2\neq 0$, $P_3\neq 0$, and $a_1=0$

We next consider

Case II:

$$a_1=0.$$

The g_i are then given by Eqs. (C3)–(C4) of Appendix C. Thus, by Eq. (A23),

$$f_1(x,t)=\frac{1}{2}\left[\frac{p+v_2q-\bar{\beta}'}{a_3}-\frac{p+v_3q-\bar{\gamma}'}{a_2}+\left(\frac{B-v_2A}{a_3}-\frac{B-v_3A}{a_2}\right)\frac{D_0'(\xi)}{D_0(\xi)}\right],$$

$$f_2(x,t)=\frac{\alpha+\alpha'}{2a_3}+a_2\frac{g_0(\xi)}{D_0(\xi)},$$

$$f_3(x,t)=-\frac{\alpha+\alpha'}{2a_2}+a_3\frac{g_0(\xi)}{D_0(\xi)}.$$

Now Q_1 is still given by (A32), whereas [see Eqs. (C6), (A19), and (A23)]

$$S_1=0,$$

$$T_1=-(\alpha+\alpha')[(a_2\beta-a_3\gamma')/2a_2a_3].$$

It follows from (A31) that

$$\frac{g_0(\xi)}{D_0(\xi)}=-\frac{\alpha+\alpha'}{2a_2a_3}\frac{a_2\beta-a_3\gamma'}{a_2\beta+a_3\gamma'}=\text{const.}$$

The distribution functions will again be of the form (A34), now with

$$\eta_1=-(B-v_3A)/a_2,$$
$$\eta_2=0,$$
$$\eta_3=0.$$

Thus, whatever the value of a_1, we conclude that for this model 1 with $P_2\neq 0$, $P_3\neq 0$, two distribution functions have to be constant, i.e., the solutions belong to class A1 [cf. Eq. (A1)].

5. Model 1 with $P_1=P_2=P_3=0$

The remaining subsections of this appendix are devoted to a study of the consistency equations of model 1, (A17), in the case of

$$P_1=a_1(B-v_1A)=0,$$
$$P_2=a_2(B-v_2A)=0, \tag{A35}$$
$$P_3=a_3(B-v_3A)=0.$$

We shall distinguish two cases:

Case III:

$$a_1=0,\quad a_2\neq 0,\quad a_3\neq 0, \tag{A36a}$$
$$B-v_2A=0, \tag{A36b}$$
$$B-v_3A=0,\quad \text{i.e.,}\quad v_2=v_3. \tag{A36c}$$

Case IV:

$$a_1=0, \tag{A37a}$$
$$a_2=0,\quad a_3\neq 0, \tag{A37b}$$
$$B-v_3A=0, \tag{A37c}$$

to be explored in subsections 6 and 7, respectively. Note that we cannot have also $a_3=0$, since then the model would no longer contain any nonlinear term.

6. Case III, $a_1=0$, $a_2\neq 0$, $a_3\neq 0$

In case III, one clearly has

$$B-v_1A\neq 0,\quad \text{i.e.,}\quad v_1\neq v_2=v_3,$$

since we do not want all three velocities to be the same. The coefficients given in Appendix C then simplify [see Eq. (C7)]. In particular, one has

$$R_1=R_2=R_3=0, \tag{A38}$$

and only two of the consistency conditions, Eqs. (A17a)–(A17d), are nontrivial:

$$Q_1S_2=Q_2S_1, \tag{A39a}$$

$$Q_1T_2=Q_2T_1. \tag{A39b}$$

With the abbreviations

$$X_2\equiv p+v_2q+\bar{\beta}, \quad X_3\equiv p+v_2q+\bar{\gamma},$$

the first of these, Eq. (A39a), leads to

$$X_2X_3=\beta'\gamma,$$

whereas the second one, Eq. (A39b), yields

$$\alpha\gamma' X_2+\alpha'\beta X_3+\alpha\beta\gamma+\alpha'\beta'\gamma'=0. \tag{A40}$$

Squaring the latter, and substituting according to the former, we obtain

$$(\alpha\gamma' X_2+\alpha'\beta X_3)^2-4\alpha\alpha'\beta\gamma'(X_2X_3)$$
$$=(\alpha\beta\gamma+\alpha'\beta'\gamma')^2-4\alpha\alpha'\beta\gamma'(\beta'\gamma)$$

or

$$\alpha\gamma' X_2-\alpha'\beta X_3=\pm(\alpha\beta\gamma-\alpha'\beta'\gamma').$$

When combining this with Eq. (A40), one can solve for X_2 and X_3. The two signs yield the solutions denoted case IIIA and case IIIB below:

Case IIIA:

$$\alpha\gamma' X_2=-\alpha'\beta'\gamma', \quad \alpha'\beta X_3=-\alpha\beta\gamma.$$

There are two possibilities:

(i) $\gamma'=0, \quad \beta\neq 0,$

$$\alpha' X_3=-\alpha\gamma, \text{ or } \alpha'(p+v_2q+\bar{\gamma})+\alpha\gamma=0;$$

(ii) $\beta=0, \quad \gamma'\neq 0,$

$$\alpha X_2=-\alpha'\beta', \text{ or } \alpha(p+v_2q+\bar{\beta})+\alpha'\beta'=0.$$

Case IIIB:

$$\alpha\gamma' X_2=-\alpha\beta\gamma, \quad \alpha'\beta X_3=-\alpha'\beta'\gamma'.$$

There are again two possibilities:

(i) $\alpha=0, \quad \alpha'\neq 0,$

$$\beta X_3=-\beta'\gamma', \text{ or } \beta(p+v_2q+\bar{\gamma})+\beta'\gamma'=0;$$

(ii) $\alpha'=0, \quad \alpha\neq 0,$

$$\gamma' X_2=-\beta\gamma, \text{ or } \gamma'(p+v_2q+\bar{\beta})+\beta\gamma=0.$$

Here, the g_i are given by Eqs. (C3) and (C4) as

$$g_1=\frac{1}{2}\left(\frac{F_2}{a_3}-\frac{F_3}{a_2}\right),$$

$$g_2=-F_1/2a_3+a_2g(x,t),$$

$$g_3=F_1/2a_2+a_3g(x,t),$$

and the three auxiliary functions F_i are given by Eqs. (A6) and (A36) as:

$$F_1=[(p+v_1q-\bar{\alpha}')D_0+(B-v_1A)D_0']e^{pt+qx},$$

$$F_2=(p+v_2q-\bar{\beta}')D_0e^{pt+qx},$$

$$F_3=(p+v_2q-\bar{\gamma}')D_0e^{pt+qx}.$$

It follows from Eq. (2.6) that

$$f_1(x,t)=\frac{1}{2}\left[\frac{p+v_2q-\bar{\beta}'}{a_3}-\frac{p+v_2q-\bar{\gamma}'}{a_2}\right],$$

$$f_2(x,t)=\frac{1}{2}\left[-\frac{p+v_1q-\bar{\alpha}'}{a_3}-\frac{B-v_1A}{a_3}\frac{D_0'(\xi)}{D_0(\xi)}\right]$$
$$+a_2\frac{g_0(\xi)}{D_0(\xi)}, \tag{A41}$$

$$f_3(x,t)=\frac{1}{2}\left[\frac{p+v_1q-\bar{\alpha}'}{a_2}+\frac{B-v_1A}{a_2}\frac{D_0'(\xi)}{D_0(\xi)}\right]$$
$$+a_3\frac{g_0(\xi)}{D_0(\xi)}.$$

By (A14), (A35), and (A38),

$$g_0(\xi)=-\frac{1}{Q_1}\left[S_1\frac{d}{d\xi}+T_1\right]D_0(\xi).$$

Here, Q_1, S_1, and T_1 are given by (C7), and

$$\frac{g_0(\xi)}{D_0(\xi)}=\frac{1}{2(a_2\beta+a_3\gamma')}\left[(B-v_1A)\left(\frac{p+v_2q-\bar{\beta}'+\beta}{a_3}\right.\right.$$
$$\left.-\frac{p+v_2q-\bar{\gamma}'+\gamma'}{a_2}\right)\frac{D_0'(\xi)}{D_0(\xi)}+(p+v_1q+\bar{\alpha})$$
$$\times\left(\frac{p+v_2q-\bar{\beta}'}{a_3}-\frac{p+v_2q-\bar{\gamma}'}{a_2}\right)$$
$$\left.+\beta\frac{p+v_1q-\bar{\alpha}'}{a_3}-\gamma'\frac{p+v_1q-\bar{\alpha}'}{a_2}\right]. \tag{A42}$$

Therefore, the distribution functions are of the form (A34). Substituting (A42) into (A41), we find

$$\eta_1=0,$$

$$\eta_2=\frac{B-v_1A}{2a_3(a_2\beta+a_3\gamma')}[-2a_3\gamma'+a_2(p+v_2q-\bar{\beta}')$$
$$-a_3(p+v_2q-\bar{\gamma}')],$$

$$\eta_3=\frac{B-v_1A}{2a_2(a_2\beta+a_3\gamma')}[2a_2\beta+a_2(p+v_2q-\bar{\beta}')$$
$$-a_3(p+v_2q-\bar{\gamma}')].$$

These solutions belong to the class A2 of Eq. (A2).

7. Case IV, $a_1=a_2=0$, $a_3\neq 0$

In case IV we have

$$A=a_3\neq 0, \quad B=a_3 v_3, \quad C=a_3\bar{\gamma}',$$

and there are three consistency conditions:

$$Q_1R_2=Q_2R_1, \tag{A43a}$$

$$Q_1S_2=Q_2S_1, \tag{A43b}$$

$$Q_1T_2=Q_2T_1. \tag{A43c}$$

The first one gives immediately

$$(\gamma+\gamma')(B-v_1A)(B-v_2A)=0,$$

and we must distinguish two subcases:

Case IVA:

$$\gamma=\gamma'=0. \tag{A44}$$

Case IVB:

$$B-v_2A=0, \text{ i.e., } v_2=v_3. \tag{A45}$$

We first consider case IVA. It follows from (A44) and Appendix C that

$$Q_1=Q_2=0,$$

and the remaining conditions, (A43b) and (A43c), are trivially satisfied.

Since for case IVA, $P_i=Q_i=0$, it follows from (A14) that we have two differential equations (with constant coefficients) for D_0. We must make sure they have at least one common solution. With the solution of the form

$$D_0=ce^{z\xi},$$

z must satisfy the two equations

$$R_1z^2+S_1z+T_1=0, \quad S_3z+T_3=0.$$

(It is convenient to consider the equation with $i=3$ instead of the one for $i=2$, since $R_3=0$.) Eliminating here z, we find the condition

$$R_1T_3^2-S_1S_3T_3+T_1S_3^2=0.$$

Substituting for these quantities according to Appendix C, we find

$$\begin{aligned}&(v_3-v_1)(v_3-v_2)[-\alpha(p+v_2q-\bar{\beta}')+\beta'(p+v_1q-\bar{\alpha}')][\alpha\beta-\beta'(\alpha+\alpha')]+\alpha\beta'\{(v_3-v_1)^2(p+v_2q-\bar{\beta}')\\&\times(p+v_2q-\bar{\beta}'+\beta)+(v_3-v_2)^2(p+v_1q-\bar{\alpha}')(p+v_1q+\bar{\alpha})-2(v_3-v_1)(v_3-v_2)[(p+v_1q+\bar{\alpha})\\&\times(p+v_2q-\bar{\beta}')+\beta(p+v_1q-\bar{\alpha}')]\}+(v_3-v_1)^2\beta'^2(\alpha+\alpha')(p+v_2q-\bar{\beta}')+(v_3-v_2)^2\alpha^2\beta(p+v_1q-\bar{\alpha}')=0.\end{aligned} \tag{A46}$$

The expression in curly brackets can be simplified. We first write it as

$$\begin{aligned}\{\cdots\}=&[(v_3-v_2)(p+v_1q-\bar{\alpha}')-(v_3-v_1)(p+v_2q-\bar{\beta}')]^2+(v_3-v_2)^2(p+v_1q-\bar{\alpha}')(\alpha+\alpha')\\&+(v_3-v_1)^2(p+v_2q-\bar{\beta}')\beta-2(v_3-v_2)(v_3-v_1)[(p+v_1q-\bar{\alpha}')\beta+(p+v_2q-\bar{\beta}')(\alpha+\alpha')],\end{aligned}$$

and then use (5.8) in the form (remember $a_1=a_2=0$)

$$p+v_3q-\bar{\gamma}'=0, \tag{A47}$$

to rewrite the square as

$$\begin{aligned}[\cdots]^2=&[(v_2-v_3)\bar{\alpha}'+(v_3-v_1)\bar{\beta}'\\&+(v_1-v_2)\bar{\gamma}']^2.\end{aligned}$$

Substituting back into (A46), we get

$$\begin{aligned}&\alpha\beta'[(v_2-v_3)\bar{\alpha}'+(v_3-v_1)\bar{\beta}'+(v_1-v_2)\bar{\gamma}']^2\\&+[(p+v_1q-\bar{\alpha}')(v_2-v_3)+(p+v_2q-\bar{\beta}')\\&\times(v_3-v_1)][\alpha(v_2-v_3)+\beta'(v_3-v_1)]\\&\times[\alpha\beta+\beta'(\alpha+\alpha')]=0.\end{aligned}$$

Invoking again (A47), we find

$$\begin{aligned}&[(v_2-v_3)\bar{\alpha}'+(v_3-v_1)\bar{\beta}'+(v_1-v_2)\bar{\gamma}']\\&\times\{\alpha\beta'[(v_2-v_3)\bar{\alpha}'+(v_3-v_1)\bar{\beta}'\\&+(v_1-v_2)\bar{\gamma}']-[(v_2-v_3)\alpha+(v_3-v_1)\beta']\\&\times[\alpha\beta+\beta'(\alpha+\alpha')]\}=0.\end{aligned}$$

This expression factorizes, so there are two possibilities:

Case IVAa:

$$(v_2-v_3)\bar{\alpha}'+(v_3-v_1)\bar{\beta}'+(v_1-v_2)\bar{\gamma}'=0,$$

Case IVAb:

$$\begin{aligned}&\alpha\beta'[(v_2-v_3)\bar{\alpha}'+(v_3-v_1)\bar{\beta}'+(v_1-v_2)\bar{\gamma}']\\&-(\alpha\beta+\alpha\beta'+\alpha'\beta')[\alpha(v_2-v_3)+\beta'(v_3-v_1)]=0.\end{aligned}$$

We next consider case IVB where we must impose

$$\gamma+\gamma'>0. \tag{A48}$$

By (A37c) and (A45),

$v_1 \neq v_2 = v_3$.

From Appendix C, it follows that [see Eq. (C14)]

$$R_1 = R_2 = R_3 = 0,$$

and there are only two nontrivial consistency conditions left. Because of (A48), Eq. (A43b) yields

$$p + v_2 q + \bar{\beta} = \beta' \gamma / (\gamma + \gamma'). \tag{A49}$$

The remaining consistency condition, (A43c), then factorizes,

$$[(\alpha + \alpha')\gamma + \alpha'\gamma'][(\beta + \beta')\gamma' + \beta\gamma] = 0.$$

We here distinguish two further subcases:

Case IVBa:

$$\alpha\gamma + \alpha'\gamma + \alpha'\gamma' = 0, \quad \text{with } \gamma + \gamma' > 0.$$

(i) $\alpha = \alpha' = 0$,

(ii) $\alpha' = \gamma = 0$.

Case IVBb:

$$\beta\gamma + \beta\gamma' + \beta'\gamma' = 0, \quad \text{with } \gamma + \gamma' > 0.$$

(i) $\beta = \beta' = 0$, (A50)

(ii) $\beta = \gamma' = 0$. (A51)

For case IV, the g_i are given by Eqs. (C10) and (C11) as

$$g_1 = F_2/a_3, \quad g_2 = -F_1/a_3, \quad g_3 = a_3 g(x,t),$$

and the three auxiliary functions F_i are given by Eqs. (A6) and (A37) as:

$$F_1 = [(p + v_1 q - \bar{\alpha}')D_0 + (B - v_1 A)D_0']e^{pt + qx},$$

$$F_2 = [(p + v_2 q - \bar{\beta}')D_0 + (B - v_2 A)D_0']e^{pt + qx},$$

$$F_3 = (p + v_2 q - \bar{\gamma}')D_0 e^{pt + qx}.$$

It follows that

$$f_1(x,t) = \frac{p + v_2 q - \bar{\beta}'}{a_3} + \frac{B - v_2 A}{a_3}\frac{D_0'(\xi)}{D_0(\xi)},$$

$$f_2(x,t) = -\frac{p + v_1 q - \bar{\alpha}'}{a_3} - \frac{B - v_1 A}{a_3}\frac{D_0'(\xi)}{D_0(\xi)}, \tag{A52}$$

$$f_3(x,t) = a_3 \frac{g_0(\xi)}{D_0(\xi)}.$$

Case IVA:

$$\gamma = \gamma' = 0.$$

By Eqs. (A44) and (C13),

$$Q_i = 0.$$

Equation (A14) yields no equation for $g_0(\xi)$. However, $D_0(\xi)$ satisfies

$$\frac{D_0'(\xi)}{D_0(\xi)} = -\frac{T_3}{S_3}$$

$$= -\frac{\alpha(p + v_2 q - \bar{\beta}') + \beta'(p + v_1 q - \bar{\alpha}')}{\alpha(B - v_2 A) - \beta'(B - v_1 A)},$$

which is a constant. Therefore,

$$\eta_1 = 0, \quad \eta_2 = 0, \quad \eta_3 = a_3.$$

The solutions belong to class A1 [cf. Eq. (A1)].

Case IVB:

$$B - v_2 A = 0, \quad \text{i.e.,} \quad v_2 = v_3. \tag{A45'}$$

By Eq. (C14):

$$Q_1 = -a_3\gamma',$$

$$R_1 = 0,$$

$$S_1 = (B - v_1 A)\frac{p + v_2 q - \bar{\beta}' + \beta}{a_3},$$

$$T_1 = (p + v_1 q + \bar{\alpha})\frac{p + v_2 q - \bar{\beta}'}{a_3} + \beta\frac{p + v_1 q - \bar{\alpha}'}{a_3},$$

and

$$\frac{g_0(\xi)}{D_0(\xi)} = \frac{1}{a_3^2\gamma'}\left[(B - v_1 A)(p + v_2 q - \bar{\beta}' + \beta)\frac{D_0'(\xi)}{D_0(\xi)} + (p + v_1 q + \bar{\alpha})(p + v_2 q - \bar{\beta}') + \beta(p + v_1 q - \bar{\alpha}')\right].$$

Substituting into Eq. (A52), and using also Eq. (A49), we obtain

$$\eta_1 = 0,$$

$$\eta_2 = -(v_2 - v_1),$$

$$\eta_3 = -(v_2 - v_1)[\beta'/(\gamma + \gamma')],$$

i.e., the solutions belong to class A2.

If, in this case IVB, $\alpha = \alpha' = 0$, two distribution functions are proportional,

$$f_3(\xi) = [\beta'/(\gamma + \gamma')]f_2(\xi).$$

Similarly, if $\beta = \gamma' = 0$,

$$f_3(\xi) = (\beta'/\gamma)f_2(\xi).$$

For cases I–IV of this Appendix, Eq. (2.1) is "trivial" in the sense that the left-hand side vanishes, either because $f_i = \text{const.}$, or because $\partial_t + v_i\partial_x = (B - v_i A)\partial_\xi$ with $B - v_i A = 0$.

APPENDIX B: MODELS WITH INITIAL DATA GIVEN BY TWO ARBITRARY FUNCTIONS

In model 1, discussed in Appendix A, the initial conditions are given in terms of *one* arbitrary function. In this appendix we shall discuss another possibility, namely when all coefficients of the differential equations (A14) vanish,

$$P_1 = P_2 = P_3 = 0, \tag{B1a}$$

$$Q_1 = Q_2 = Q_3 = 0, \tag{B1b}$$

$$R_1=R_2=R_3=0, \tag{B1c}$$

$$S_1=S_2=S_3=0, \tag{B1d}$$

$$T_1=T_2=T_3=0. \tag{B1e}$$

There is then no relation between g_0 and D_0, and the initial ($t=0$) condition can be given in terms of *two* arbitrary functions. In order to satisfy Eq. (B1a) without having all three velocities the same, at least one of the a_i must vanish. Accordingly, we distinguish two cases, depending on whether one or two of the a_i's vanish.

1. Case V, $a_1=0$, $a_2\neq 0$, $a_3\neq 0$

We here consider the case

$$a_1=0, \quad a_2\neq 0, \quad a_3\neq 0.$$

By Eq. (B1a) and (A12a),

$$B-v_2A=B-v_3A=0. \tag{B2}$$

Therefore,

$$v_1\neq v_2=v_3. \tag{B3}$$

For the remaining conditions (B1b)–(B1e), we consider those with $i=2$ and 3. Because of Eq. (B3), (B1c) is automatically satisfied, and the conditions on the Q_i and the S_i are:

$$Q_2=0: \quad a_2(p+v_2q+\bar{\beta})=a_3\gamma,$$

$$Q_3=0: \quad a_3(p+v_2q+\bar{\gamma})=a_2\beta',$$

$$S_2=0: \quad a_2(p+v_2q+\bar{\beta})=-a_3\gamma,$$

$$S_3=0: \quad a_3(p+v_2q+\bar{\gamma})=-a_2\beta',$$

where we have also used (B3). It follows that

$$p+v_2q+\bar{\beta}=p+v_2q+\bar{\gamma}=0, \tag{B4}$$

$$\beta'=\gamma=0. \tag{B5}$$

Substituting these results (B4) and (B5) into Eq. (5.8), invoking also (2.11), we find

$$a_2\beta+a_3\gamma'=0. \tag{B6}$$

We next make use of the results (B4) and (B5) in the conditions $T_2=T_3=0$:

$$T_2=0: \quad \alpha'(a_2\beta-a_3\gamma')=0,$$

$$T_3=0: \quad \alpha(a_2\beta-a_3\gamma')=0.$$

Because of (B6) and the requirement that $f_1(x,t)\neq 0$ [see Eq. (B9)], there is only one solution,

$$\alpha=\alpha'=0. \tag{B7}$$

The g_i are given by Eqs. (C3) and (C4) as

$$g_1=\frac{1}{2}\left(\frac{F_2}{a_3}-\frac{F_3}{a_2}\right),$$

$$g_2=-F_1/2a_3+a_2g(x,t), \tag{B8}$$

$$g_3=F_1/2a_2+a_3g(x,t),$$

and the auxiliary functions F_i are given by Eqs. (A6), (B2), (B4), and (B5) as:

$$F_1=[(p+v_1q-\bar{\alpha}')D_0+(B-v_1A)D_0']e^{pt+qx},$$

$$F_2=-\beta D_0e^{pt+qx},$$

$$F_3=-\gamma' D_0e^{pt+qx}.$$

It follows from Eqs. (2.6), (B8) and (B7) that

$$f_1(x,t)=-\frac{1}{2}\left(\frac{\beta}{a_3}-\frac{\gamma'}{a_2}\right),$$

$$f_2(x,t)=\frac{1}{2}\left[-\frac{p+v_1q+\bar{\alpha}}{a_3}-\frac{B-v_1A}{a_3}\frac{D_0'(\xi)}{D_0(\xi)}\right]+a_2\frac{g_0(\xi)}{D_0(\xi)}, \tag{B9}$$

$$f_3(x,t)=\frac{1}{2}\left[\frac{p+v_1q+\bar{\alpha}}{a_2}+\frac{B-v_1A}{a_2}\frac{D_0'(\xi)}{D_0(\xi)}\right]+a_3\frac{g_0(\xi)}{D_0(\xi)}.$$

Therefore, the distribution functions depend on *two* "arbitrary" functions $\phi(\xi)$ and $\psi(\xi)$ as follows:

$$f_i(x,t)=f_{i0}+\eta_i\phi(\xi)+\sigma_i\psi(\xi),$$

i.e., the solutions belong to the class A2 of Eq. (A2).

Using (B6), we obtain from (B9),

$$f_{10}=-\frac{1}{2a_2a_3}(a_2\beta-a_3\gamma')=-\frac{\beta}{a_3}=\frac{\gamma'}{a_2},$$

$$f_{20}=-(1/2a_3)(p+v_1q+\bar{\alpha}),$$

$$f_{30}=(1/2a_2)(p+v_1q+\bar{\alpha}),$$

$$\eta_1=0,$$

$$\eta_2=-(B-v_1A)/2a_3,$$

$$\eta_3=(B-v_1A)/2a_2,$$

$$\sigma_1=0, \quad \sigma_2=a_2, \quad \sigma_3=a_3.$$

These only give a nontrivial solution when

$$\beta\neq 0, \quad \gamma'\neq 0, \quad a_2>0, \quad a_3<0, \quad a_2\beta=-a_3\gamma'.$$

2. Case VI, $a_1=a_2=0$, $a_3\neq 0$

We next consider

$$a_1=a_2=0, \quad a_3\neq 0.$$

The conditions (cf. Appendix C)

$$R_1=R_2=0,$$

then require

$$B-v_2A=0. \tag{B10}$$

By a choice of labeling, we take $B-v_1A\neq 0$, i.e., $v_1\neq v_2$. It also follows from (C8) that $B-v_3A=0$. Thus, like for case V,

$v_1 \neq v_2 = v_3$.

Next, we note that $Q_1 = Q_2 = 0$ require

$$\gamma = \gamma' = 0. \tag{B11}$$

Furthermore, invoking (B10), $S_2 = S_3 = 0$ require

$$p + v_2 q + \bar{\beta} = 0, \tag{B12}$$

$$\beta' = 0. \tag{B13}$$

Incidentally, we also have $Q_3 = 0$, i.e.,

$$p + v_2 q + \bar{\gamma} = 0. \tag{B14}$$

Hence

$$\bar{\beta} = \bar{\gamma}.$$

Finally, the requirements $T_2 = T_3 = 0$ imply

$$\alpha'(p + v_2 q - \bar{\beta}') = 0, \quad \alpha(p + v_2 q - \bar{\beta}') = 0.$$

Taken together with (B12) and (B13), these imply

$$\alpha'\beta = \alpha\beta = 0.$$

There are two solutions:

$$\text{(i)} \quad \alpha = \alpha' = 0,$$

$$\text{(ii)} \quad \beta = 0. \tag{B15}$$

Here, the g_i are given by

$$g_1 = F_2/a_3, \quad g_2 = -F_1/a_3, \quad g_3 = a_3 g(x,t).$$

By (B3), and (B10)–(B14),

$$F_1 = [(p + v_1 q - \bar{\alpha}')D_0 + (B - v_1 A)D_0']e^{pt + qx},$$

$$F_2 = -\beta D_0 e^{pt + qx},$$

$$F_3 = 0,$$

and thus

$$f_1(x,t) = -\beta/a_3,$$

$$f_2(x,t) = -\frac{p + v_1 q - \bar{\alpha}'}{a_3} - \frac{B - v_1 A}{a_3}\frac{D_0'(\xi)}{D_0(\xi)},$$

$$f_3(x,t) = a_3[g_0(\xi)/D_0(\xi)],$$

i.e., the solutions are again of class A2. The second solution of Eq. (B15) is not allowed, as it would lead to $f_1(x,t) \equiv 0$.

For both these cases V and VI of this appendix, Eq. (2.1) is "trivial" in the sense that the left-hand side vanishes, either because $f_i = \text{const}$, or because $\partial_t + v_i \partial_x = (B - v_i A)\partial_\xi$ with $B - v_i A = 0$.

APPENDIX C: SPECIAL CASES OF EQS. (A11)

1. $a_1 = 0$, $a_2 \neq 0$, $a_3 \neq 0$

When $a_1 = 0$, one has

$$A = a_2 + a_3, \quad B = a_2 v_2 + a_3 v_3, \quad C = a_2 \bar{\beta}' + a_3 \bar{\gamma}', \tag{C1}$$

and Eq. (A5) takes the form

$$a_2 F_2 + a_3 F_3 = 0. \tag{C2}$$

Now, Eq. (A7) is solved by

$$g_1 = g_{10},$$

$$g_2 = g_{20} + a_2 g(x,t), \tag{C3}$$

$$g_3 = g_{30} + a_3 g(x,t),$$

with

$$g_{10} = \frac{1}{2}\left(\frac{F_2}{a_3} - \frac{F_3}{a_2}\right),$$

$$g_{20} = -\frac{1}{2}\frac{F_1}{a_3}, \tag{C4}$$

$$g_{30} = \frac{1}{2}\frac{F_1}{a_2}.$$

We note that the latter can be formally obtained from (A8) by the substitutions

$$\frac{1}{a_1} \to 0, \quad \frac{1}{3} \to \frac{1}{2}. \tag{C5}$$

We can then obtain the coefficients of Eq. (A11) from Eqs. (A12a)–(A12e) and cyclic permutations thereof, applying the substitution (C5) to those factors that originate from the g_{i0}. The result is as follows:

$$P_1 = 0,$$

$$P_2 = a_2(B - v_2 A) = -a_2 a_3 (v_2 - v_3),$$

$$P_3 = a_3(B - v_3 A) = a_2 a_3 (v_2 - v_3),$$

$$Q_1 = -a_2\beta - a_3\gamma',$$

$$Q_2 = a_2(p + v_2 q + \bar{\beta}) - a_3\gamma,$$

$$Q_3 = a_3(p + v_3 q + \bar{\gamma}) - a_2\beta',$$

$$R_1 = \frac{1}{2}(B - v_1 A)\left(\frac{B - v_2 A}{a_3} - \frac{B - v_3 A}{a_2}\right)$$

$$= -(B - v_1 A)(v_2 - v_3),$$

$$R_2 = \frac{1}{2}(B - v_2 A)\left(-\frac{B - v_1 A}{a_3}\right) = \frac{1}{2}(B - v_1 A)(v_2 - v_3),$$

$$R_3 = \frac{1}{2}(B - v_3 A)\left(\frac{B - v_1 A}{a_2}\right) = \frac{1}{2}(B - v_1 A)(v_2 - v_3), \tag{C6}$$

$$S_1 = \frac{1}{2}\Bigg[(B - v_1 A) \times \left(\frac{p + v_2 q - \bar{\beta}' + \beta}{a_3} - \frac{p + v_3 q - \bar{\gamma}' + \gamma'}{a_2}\right) + (p + v_1 q + \bar{\alpha})\left(\frac{B - v_2 A}{a_3} - \frac{B - v_3 A}{a_2}\right)\Bigg],$$

$$S_2 = \frac{1}{2}\Bigg[(B - v_1 A)\left(-\frac{p + v_2 q + \bar{\beta}}{a_3} - \frac{\gamma}{a_2}\right) - (B - v_2 A) \times \frac{p + v_1 q - \bar{\alpha}' + \alpha'}{a_3} + (B - v_3 A)\frac{\alpha'}{a_2}\Bigg],$$

$$S_3=\frac{1}{2}\left[(B-v_1A)\left(\frac{\beta'}{a_3}+\frac{p+v_3q+\bar{\gamma}}{a_2}\right)-(B-v_2A)\frac{\alpha}{a_3}\right.$$
$$\left.+(B-v_3A)\frac{p+v_1q-\bar{\alpha}'+\alpha}{a_2}\right],$$

$$T_1=\frac{1}{2}\left[(p+v_1q+\bar{\alpha})\left(\frac{p+v_2q-\bar{\beta}'}{a_3}-\frac{p+v_3q-\bar{\gamma}'}{a_2}\right)\right.$$
$$\left.+\beta\frac{p+v_1q-\bar{\alpha}'}{a_3}-\gamma'\frac{p+v_1q-\bar{\alpha}'}{a_2}\right],$$

$$T_2=\frac{1}{2}\left[(p+v_2q+\bar{\beta})\left(-\frac{p+v_1q-\bar{\alpha}'}{a_3}\right)\right.$$
$$-\gamma\frac{p+v_1q-\bar{\alpha}'}{a_2}$$
$$\left.-\alpha'\left(\frac{p+v_2q-\bar{\beta}'}{a_3}-\frac{p+v_3q-\bar{\gamma}'}{a_2}\right)\right],$$

$$T_3=\frac{1}{2}\left[(p+v_3q+\bar{\gamma})\frac{p+v_1q-\bar{\alpha}'}{a_2}-\alpha\left(\frac{p+v_2q-\bar{\beta}'}{a_3}\right.\right.$$
$$\left.\left.-\frac{p+v_3q-\bar{\gamma}'}{a_2}\right)-\beta'\left(-\frac{p+v_1q-\bar{\alpha}'}{a_3}\right)\right].$$

Some of the cases discussed in Appendices A and B require $B-v_2A=B-v_3A=0$.

The above expressions then simplify further:

$$P_1=P_2=P_3=0,$$

$$Q_1=-a_2\beta-a_3\gamma',$$

$$Q_2=a_2(p+v_2q+\bar{\beta})-a_3\gamma,$$

$$Q_3=a_3(p+v_2q+\bar{\gamma})-a_2\beta',$$

$$R_1=R_2=R_3=0,$$

$$S_1=\frac{1}{2}(B-v_1A)\left(\frac{p+v_2q-\bar{\beta}'+\beta}{a_3}\right.$$
$$\left.-\frac{p+v_2q-\bar{\gamma}'+\gamma'}{a_2}\right), \tag{C7}$$

$$S_2=\frac{1}{2}(B-v_1A)\left(-\frac{p+v_2q+\bar{\beta}}{a_3}-\frac{\gamma}{a_2}\right),$$

$$S_3=\frac{1}{2}(B-v_1A)\left(\frac{\beta'}{a_3}+\frac{p+v_2q+\bar{\gamma}}{a_2}\right),$$

$$T_1=\frac{1}{2}\left[(p+v_1q+\bar{\alpha})\left(\frac{p+v_2q-\bar{\beta}'}{a_3}-\frac{p+v_2q-\bar{\gamma}'}{a_2}\right)\right.$$
$$\left.+\beta\frac{p+v_1q-\bar{\alpha}'}{a_3}-\gamma'\frac{p+v_1q-\bar{\alpha}'}{a_2}\right],$$

$$T_2=\frac{1}{2}\left[(p+v_2q+\bar{\beta})\left(-\frac{p+v_1q-\bar{\alpha}'}{a_3}\right)\right.$$
$$-\gamma\frac{p+v_1q-\bar{\alpha}'}{a_2}$$
$$\left.-\alpha'\left(\frac{p+v_2q-\bar{\beta}'}{a_3}-\frac{p+v_2q-\bar{\gamma}'}{a_2}\right)\right],$$

$$T_3=\frac{1}{2}\left[(p+v_2q+\bar{\gamma})\frac{p+v_1q-\bar{\alpha}'}{a_2}-\alpha\left(\frac{p+v_2q-\bar{\beta}'}{a_3}\right.\right.$$
$$\left.\left.-\frac{p+v_2q-\bar{\gamma}'}{a_2}\right)-\beta'\left(-\frac{p+v_1q-\bar{\alpha}'}{a_3}\right)\right].$$

2. $a_1=a_2=0$, $a_3\neq 0$

When $a_1=a_2=0$, one has

$$A=a_3,\quad B=a_3v_3,\quad C=a_3\bar{\gamma}', \tag{C8}$$

and Eq. (A5) reduces to

$$a_3F_3=0. \tag{C9}$$

Furthermore, Eq. (A7) is solved by

$$g_1=g_{10},\quad g_2=g_{20},\quad g_3=g_{30}+a_3g(x,t), \tag{C10}$$

with

$$g_{10}=F_2/a_3,\quad g_{20}=-F_1/a_3,\quad g_{30}=0. \tag{C11}$$

These latter relations can be formally obtained from (A8) by the substitutions

$$1/a_1\to 0,\quad 1/a_2\to 0,\quad \tfrac{1}{3}\to 1, \tag{C12}$$

and we obtain

$$P_1=P_2=P_3=0,$$

$$Q_1=-a_3\gamma',$$

$$Q_2=-a_3\gamma,$$

$$Q_3=a_3(p+v_3q+\bar{\gamma})=a_3(\gamma+\gamma'),$$

$$R_1=(B-v_1A)[(B-v_2A)/a_3],$$

$$R_2=-(B-v_1A)[(B-v_2A)/a_3],$$

$$R_3=0, \tag{C13}$$

$$S_1=(B-v_1A)\frac{p+v_2q-\bar{\beta}'+\beta}{a_3}$$
$$+(B-v_2A)\frac{p+v_1q+\bar{\alpha}}{a_3},$$

$$S_2=-(B-v_1A)\frac{p+v_2q+\bar{\beta}}{a_3}$$
$$-(B-v_2A)\frac{p+v_1q-\bar{\alpha}'+\alpha'}{a_3},$$

$$S_3=(B-v_1A)(\beta'/a_3)-(B-v_2A)(\alpha/a_3),$$

$$T_1=(p+v_1q+\bar{\alpha})\frac{p+v_2q-\bar{\beta}'}{a_3}+\beta\frac{p+v_1q-\bar{\alpha}'}{a_3},$$

$$T_2=-(p+v_2q+\bar{\beta})\frac{p+v_1q-\bar{\alpha}'}{a_3}-\alpha'\frac{p+v_2q-\bar{\beta}'}{a_3},$$

$$T_3=-\alpha\frac{p+v_2q-\bar{\beta}'}{a_3}-\beta'\frac{p+v_1q-\bar{\alpha}'}{a_3}.$$

Some of the cases of Appendices A and B require $B-v_2A=0$. The above quantities then simplify further:

$$P_1=P_2=P_3=0,$$

$$Q_1=-a_3\gamma',$$

$$Q_2=-a_3\gamma,$$

$$Q_3=a_3(p+v_3q+\bar{\gamma})=a_3(\gamma+\gamma'),$$

$$R_1=R_2=R_3=0,$$

$$S_1=(B-v_1A)\frac{p+v_2q-\bar{\beta}'+\beta}{a_3}, \tag{C14}$$

$$S_2=-(B-v_1A)\frac{p+v_2q+\bar{\beta}}{a_3},$$

$$S_3=(B-v_1A)(\beta'/a_3),$$

$$T_1=(p+v_1q+\bar{\alpha})\frac{p+v_2q-\bar{\beta}'}{a_3}+\beta\frac{p+v_1q-\bar{\alpha}'}{a_3},$$

$$T_2=-(p+v_2q+\bar{\beta})\frac{p+v_1q-\bar{\alpha}'}{a_3}-\alpha'\frac{p+v_2q-\bar{\beta}'}{a_3},$$

$$T_3=-\alpha\frac{p+v_2q-\bar{\beta}'}{a_3}-\beta'\frac{p+v_1q-\bar{\alpha}'}{a_3}.$$

APPENDIX D: ASYMPTOTIC BEHAVIOR OF $\bar{H}(\zeta,t)$

Starting with the (matrix) retarded Green's function $G(x,t)$ defined by (7.10)–(7.12), we have identified and separated out the δ-function parts, as given by (7.46). After removing the δ-function parts, the remainder, called $H(x,t)$, still contains finite discontinuities. For the purposes of the numerical computation, it is essential to know what these finite discontinuities are. It is the purpose of this appendix to provide this information.

It is seen from (7.32), (7.34), and (7.36)–(7.38) that these discontinuities occur only in the off-diagonal elements of $H(x,t)$. It is therefore sufficient to give the leading ζ^{-1} terms for these elements of $\bar{H}(\zeta,t)$. They are

$$\begin{aligned}
\bar{H}_{12}(\zeta,t)&\sim i\zeta^{-1}[\beta/(v_1-v_2)][e^{-(iv_1\zeta+\alpha+\alpha')t}\\
&\quad-e^{-(iv_2\zeta+\beta+\beta')t}],\\
\bar{H}_{23}(\zeta,t)&\sim i\zeta^{-1}[\gamma/(v_2-v_3)][e^{-(iv_2\zeta+\beta+\beta')t}\\
&\quad-e^{-(iv_3\zeta+\gamma+\gamma')t}],\\
\bar{H}_{31}(\zeta,t)&\sim i\zeta^{-1}[\alpha/(v_3-v_1)][e^{-(iv_3\zeta+\gamma+\gamma')t}\\
&\quad-e^{-(iv_1\zeta+\alpha+\alpha')t}],\\
\bar{H}_{21}(\zeta,t)&\sim i\zeta^{-1}[\alpha'/(v_1-v_2)][e^{-(iv_1\zeta+\alpha+\alpha')t}\\
&\quad-e^{-(iv_2\zeta+\beta+\beta')t}],\\
\bar{H}_{32}(\zeta,t)&\sim i\zeta^{-1}[\beta'/(v_2-v_3)][e^{-(iv_2\zeta+\beta+\beta')t}\\
&\quad-e^{-(iv_3\zeta+\gamma+\gamma')t}],\\
\bar{H}_{13}(\zeta,t)&\sim i\zeta^{-1}[\gamma'/(v_3-v_1)][e^{-(iv_3\zeta+\gamma+\gamma')t}\\
&\quad-e^{-(iv_1\zeta+\alpha+\alpha')t}].
\end{aligned} \tag{D1}$$

The leading terms of the diagonal elements $\bar{H}_{jj}(\zeta,t)$ are all of order ζ^{-2} for large ζ.

[1] M. Krook and T. T. Wu, Phys. Rev. Lett. **36**, 1107 (1976).
[2] A. V. Bobylev, Sov. Phys. Dokl. **20**, 820 (1976).
[3] K. Olaussen, Phys. Rev. A **25**, 3393 (1982).
[4] H. P. McKean, Arch. Ration. Mech. Anal. **21**, 343 (1966); S. Harris, J. Math. Phys. **8**, 2407 (1967).
[5] T. Carleman, *Problèmes Mathématiques dans la Théorie Cinétique des Gaz* (Almqvist & Wiksells, Uppsala, 1957); I. I. Kolodner, Cenn. Mat. Pura Appl. (Ser. 4) **63**, 11 (1963); J. E. Broadwell, Phys. Fluids **7**, 1243 (1964); H. P. McKean, J. Comb. Theory **2**, 358 (1967); J. Hardy, Y. Pomeau, and O. de Pazzis, J. Math. Phys. **14**, 1746 (1973); M. H. Ernst and E. M. Hendriks, Phys. Lett. A **70**, 183 (1979); R. Illner, Math. Meth. Appl. Sci. **1**, 187 (1979); Comm. Math. Phys. **93**, 341 (1984); G. Spiga, Phys. Fluids **27**, 2599 (1984); V. C. Boffi and G. Spiga, Phys. Rev. A **29**, 782 (1984); T. Platkowski, J. Mec. Theor. Appl. **4**, 555 (1985); V. C. Boffi and T. F. Nonnenmacher, Nuovo Cimento B **85**, 165 (1985); T. F. Nonnenmacher, Phys. Rev. A **34**, 1599 (1986); G. Dukek and T. F. Nonnenmacher, Physica A **135**, 167 (1986); A. Santos and J. J. Brey, Phys. Lett. A **123**, 268 (1987); B. I. Sadovnikov and N. G. Inozemtseva, Phys. Lett. A **124**, 120 (1987).
[6] H. Cornille, J. Math. Phys. **27**, 1373 (1986); **28**, 1567 (1987); Phys. Lett. A **125**, 253 (1987); J. Stat. Phys. **48**, 789 (1987).
[7] Th. W. Ruijgrok and T. T. Wu, Physica A **113**, 401 (1982).
[8] J. H. Christenson, J. W. Cronin, V. L. Fitch, and R. Turlay, Phys. Rev. Lett. **13**, 138 (1964).
[9] E. C. Stueckelberg, Helv. Phys. Acta **25**, 577 (1952); J. S. Thomson, Phys. Rev. **91**, 1263 (1953); M. Inagaki, G. Wanders, and C. Piron, Helv. Phys. Acta **27**, 71 (1954); A. Aharony, in *Modern Developments in Thermodynamics*, edited B. Gal-Or (Wiley, New York, 1974); C. N. Yang and C. P. Yang, Trans. N. Y. Acad. Sci. Ser. II **40**, 267 (1980).
[10] F. R. Gantmacher, *The Theory of Matrices* (Chelsea, New York, 1960), Vol. II, p. 12.
[11] P. Osland and T. T. Wu (to be published).

008012JMP = 008012JMP

Equilibrium properties of a multi-urn model

B. Gaveau
Department of Mathematics, University of Paris VI, 4 pl. Jussieu, 75252 Paris Cedex 05, France

L. S. Schulman
Physics Department, Clarkson University, Potsdam, New York 13699-5820

(Received 13 March 1990; accepted for publication 18 July 1990)

Equilibrium properties of the following model are studied: A system is formed of urns (or droplets, or clusters), each containing a number of atoms. That number is governed by an occupation-dependent energy and a chemical potential per atom. An effective interaction between clusters arises from a global constraint on total cluster number and occupation. The system phase diagram exhibits various kinds of transitions. The microscopic structure of the thermodynamic state is also examined (cluster occupation probabilities, etc.) and surprising simplifications relating it to the single-urn model are found. In general, the system is intermediate in its complexity between the single-urn model and the Fisher–Felderhof droplet model. The ultimate purpose in defining this model is to study its dynamical properties and, in particular, its metastable phases. The cluster size distributions are anticipated to be of particular importance for this issue.

I. INTRODUCTION

The goal of this article is to develop a model on which to study nonequilibrium behavior. Such an objective has implicit conflicts. On the one hand, solvability of the dynamics imposes severe constraints of simplicity while yielding to these constraints results in models that fail to capture features of realistic systems. The model we shall propose is a version of the Fisher–Felderhof droplet[1,2] model, but a version that has been significantly simplified to allow us to gain dynamical information.

In the Fisher–Felderhof model, one has droplets moving about and able to combine or break up through contact interactions; other interdroplet forces are absent. Although the latter feature is a significant departure from reality, the model still has physical structure and, in particular, retains space-time translation degrees of freedom for the droplets. In our model, we retain some hints of that space-time structure. The model consists of a collection of urns (or "clusters" or "droplets;" the terms will be used interchangably) of the sort that we have previously used in the study of metastability.[3,4] However, that earlier model—which *was* dynamically tractable—is now enriched by allowing a large number of them. The clusters interact by means of a constraint that is analogous to that used in the spherical model, and the spatial structure manifests itself by our allowing an empty site for each cluster and a reservoir.

In the second section we define the model and the partition function. The thermodynamic properties and the phase diagrams are developed in Secs. III and IV. In order to study nonthermodynamic properties, we need a rather precise estimate of the partition function. The necessary mathematical tools are given in Secs. V and VI, and are applied in Secs. VII to X. With them we compute the average number of clusters of a given size, the average number of particles in the reservoir, the total number of clusters, the length of a cluster, and other properties. We see, in particular, that in the condensed phase there is basically only one cluster. Section XI discusses the meaning of the equation of state in the gaseous phase and gives an alternative (although nonrigorous) derivation of the equation of state using a mean field type argument.

II. PARTITION FUNCTION

The system has total size L, which consists of a reservoir and clusters. A configuration is determined by the set of numbers N_n of clusters of size n ($= 1,2,...$) and by the number P of particles in the reservoir. These quantities satisfy the constraint

$$\sum_n nN_n + \sum_n N_n + P = L. \tag{2.1}$$

The second sum corresponds to having one "vacuum" site per cluster in the total "geometry" L.

One can think of our system and the constraint (2.1) as a kind of mean field or spherical model of a one-dimensional lattice droplet model. Imagine L sites, which may be occupied or unoccupied, with occupied contiguous sites forming clusters (or droplets). A state of such a system would be characterized by the sequence of cluster sizes as well as the sizes of the gaps between clusters. Our model represents a simplification of that system in the following way: The distribution of gap sizes is ignored so long as any pair of clusters is separated by at least one space [these spaces are what give rise to $\Sigma_n N_n$ in Eq. (2.1)]. The energy of a configuration $\{N_n,P\}$ does not depend on the order of the $\{N_n\}$ although distinct sequences are counted separately. For example, the specification $N_1 = 1$, $N_3 = 1$, $N_5 = 1$, others zero, corresponds to several possible states: A cluster of size 5, followed by a 3, followed by a 1; or (1,3,5) or (1,5,3) or (5,1,3) or (3,1,5) or (3,5,1). However, $N_1 = 2$ (others zero) represents only a single possibility: (1,1). Thus our atoms or balls

 0022-2488/90/123030-11$03.00

are not distinguishable. This picture will be further elaborated in our intended future publication when we will select our dynamical rules for the clusters by working up from simpler dynamical rules for the balls, together with the indistinguishability of the balls.

We assign an energy $E(n)$ to a cluster of size n, for example, the droplet model expression:[1-4]

$$E(n) = \log n - \mu n$$

and a chemical potential ν per particle in the reservoir. The partition function is therefore

$$Z_L = \sum_{\{N_n,P\}}{}' \frac{(\Sigma_n N_n)!}{\Pi_n N_n!}\left(\prod_n e^{-\beta E(n)N_n}\right)e^{-\beta\nu P}, \tag{2.2}$$

where the prime on the summation symbol indicates that only sets $\{N_n,P\}$ satisfying Eq. (2.1) are used. Then we obtain the following generating function:

$$\sum_{L>0} x^L Z_L = (1-e^{-\beta\nu}x)^{-1}\left(1 - x\sum_{n=1}^{\infty}\frac{1}{n^\beta}e^{\beta\mu n}x^n\right)^{-1}. \tag{2.3}$$

For now we take $\nu \neq 0$, so as to compute various quantities, but at a later stage we will set $\nu = 0$. We also define

$$\phi_{\beta,\mu}(x) = \sum_{n=1}^{\infty}\frac{1}{n^\beta}e^{\beta\mu n}x^n \quad \text{(urn partition function)}. \tag{2.4}$$

III. THERMODYNAMICS

The thermodynamics of this model is controlled by the large L behavior of

$$\log Z_L/L = \log(Z_L^{1/L}),$$

which is precisely the logarithm of the inverse of the radius of convergence of the generating function (2.3). If we set $\nu = 0$ in (2.3), then the first singularity of (2.3) in x is, apart from $x = 1$, (i) either the singularity of $\phi_{\beta,\mu}(x)$ or (ii) the zero of $1 - x\phi_{\beta,\mu}(x) = 0$.

A. $\beta \leqslant 1$ (gas phase)

The function $\phi_{\beta,\mu}(x)$ has an infinity at $x = e^{-\beta\mu}$ which for $\mu > 0$ is less than 1. Then the root of the equation

$$1/x = \phi_{\beta,\mu}(x) \tag{3.1}$$

is a certain point x_0, the intersection of $1/x$ with $\phi_{\beta,\mu}(x)$, which is less than $e^{-\beta\mu}$ (see Fig. 1). This intersection always exists for $\beta \leqslant 1$ and depends smoothly on everything. Thinking of our model as a gas of clusters we define the pressure to be

$$p \equiv \frac{1}{\beta}\lim_{L\to\infty}\frac{1}{L}\log Z_L = -\frac{1}{\beta}\log x_0. \tag{3.2}$$

Since $\phi_{\beta,\mu}(x_0) < \infty$, we have

$$\sum_{n>1}\frac{1}{n^\beta}(e^{\beta\mu}x_0)^n < \infty,$$

so that $e^{\beta\mu}x_0 < 1$. This implies that if $\mu \to +\infty$, $x_0 \to 0^+$, and when $\mu \to -\infty$, $x_0 \to +\infty$.

B. $\beta > 1$ (condensed phase or gas phase, depending on μ)

For $\beta > 1$, $\phi_{\beta,\mu}(x)$ becomes singular but stays finite and equal to $\Sigma(1/n^\beta)$ for $x = e^{-\beta\mu}$. We define $\mu_c(\beta)$ by the equation

$$e^{\beta\mu_c(\beta)} = \sum_{n>1}\frac{1}{n^\beta} \tag{3.3}$$

and we have the following three cases.

(a) Gas phase $0 < \mu < \mu_c(\beta)$.

In this case the curve $\phi_{\beta,\mu}(x)$ cuts the line $x = e^{-\beta\mu}$ at $\Sigma_{n>1}(1/n^\beta) > e^{\beta\mu}$, so that the two curves $1/x$ and $\phi_{\beta,\mu}(x)$ intersect each other at a point $x_0 < e^{-\mu\beta}$ as in Fig. 2. Then

$$\beta p = \lim(\log Z_L/L) = -\log x_0$$

and βp depends smoothly on everything in this domain.

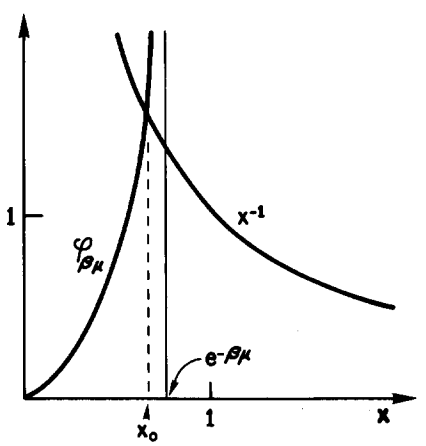

FIG. 1. The "urn partition function," $\phi_{\beta,\mu}(x)$ for the case $\beta < 1$. As $x \uparrow e^{-\beta\mu}$, $\mu \to \infty$ so that $\phi_{\beta,\mu}(x)$ meets the curve $1/x$ to the left of $e^{-\beta\mu}$, which for $\mu > 0$ is less than 1.

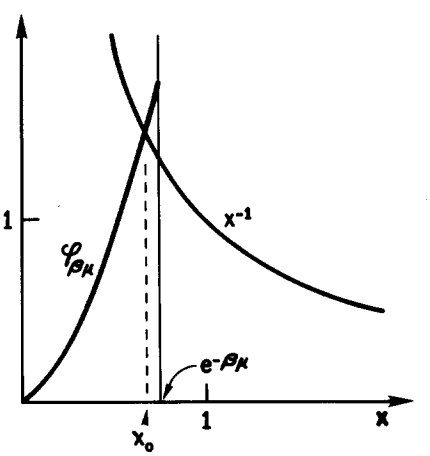

FIG. 2. When $\beta > 1$ and $\mu < \mu_c(\beta)$, $\phi_{\beta,\mu}(x)$ remains finite for $x \uparrow e^{-\beta\mu}$.

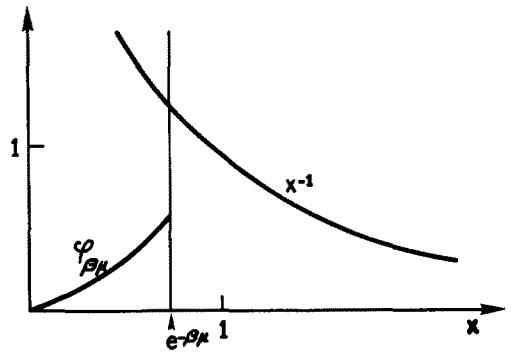

FIG. 3. For $\beta > 1$ and $\mu > \mu_c(\beta)$ the curves $\phi_{\beta,\mu}(x)$ and $1/x$ do not intersect and the singularity in x occurs at $e^{-\beta\mu}$. As a consequence, $p = \mu$ and $\rho = 1$.

and βp depends smoothly on everything in this domain.

(b) Condensed phase $\mu > \mu_c(\beta)$

In this case $\phi_{\beta,\mu}(x)$ cuts the line $x = e^{-\beta\mu}$ at $\Sigma(1/n^\beta) < e^{\beta\mu}$, so that the two curves $1/x$ and $\phi_{\beta,\mu}(x)$ do not intersect (see Fig. 3). The singularity of $\Sigma x^L Z_L$ is then at $x_0 = e^{-\beta\mu}$ and the pressure is

$$p = -(1/\beta)\log x_0 = \mu. \tag{3.4}$$

The density is $\rho = \partial p/\partial\mu = 1$, so that we are in a condensed phase.

(c) Intermediate case $\mu = \mu_c(\beta)$

In this case $\phi_{\mu,\beta}$ and $1/x$ intersect at $\Sigma_{n>1}(1/n^\beta)$ for $x_0 = e^{-\beta\mu}$, as in Fig. 4.

IV. PHASE DIAGRAMS

In the noncondensed or gas phase

$$\frac{1}{x_0} = \sum_{n>1} \frac{(e^{\beta\mu})^n x_0^n}{n^\beta}. \tag{4.1}$$

It follows that the density is

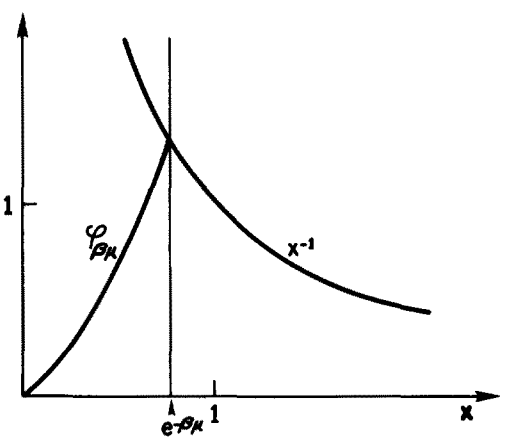

FIG. 4. Intermediate case, $\mu = \mu_c(\beta)$. The singularities coincide.

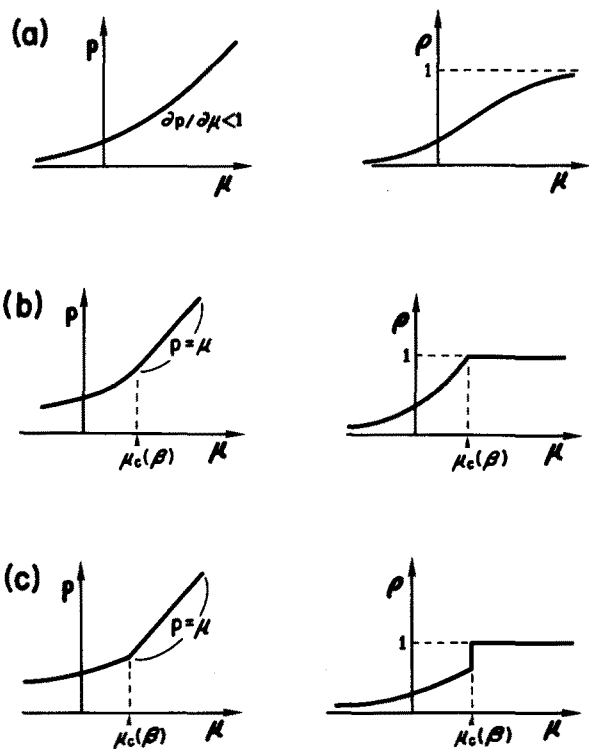

FIG. 5. Pressure (p) and density (ρ) variation for various ranges of β. (a) $\beta \leqslant 1$. $\partial p/\partial\mu < 1$ for all μ. (b) $1 < \beta \leqslant 2$. Slope is continuous, next derivative is not. (c) $\beta > 2$. Slope is discontinuous.

$$\rho = \frac{\partial p}{\partial \mu} = -\frac{1}{\beta}\frac{\partial \log x_0}{\partial \mu} = \frac{\Sigma(e^{\beta\mu})^n (n x_0^n/n^\beta)}{1/x_0 + \Sigma(e^{\beta\mu})^n (n x_0^n/n^\beta)}. \tag{4.2}$$

This equation is obtained by computing the derivative with respect to μ of Eq. (4.1) and it is thus obvious that $\rho < 1$. In Eq. (4.2) we let $\mu \to \mu_c^-$ and because $x_0 = e^{-\mu\beta_c}$ at $\mu = \mu_c$ we obtain

$$\left.\frac{\partial p}{\partial \mu}\right|_{\mu_c^-} = \frac{\Sigma(1/n^{\beta-1})}{\Sigma(1/n^\beta) + \Sigma(1/n^{\beta-1})}. \tag{4.3}$$

Now if $1 < \beta \leqslant 2$, this is 1 because $\Sigma(1/n^{\beta-1}) = \infty$ for $1 < \beta \leqslant 2$; if $\beta > 2$, this is a certain fixed finite number (smaller than 1). If $\mu \to +\infty$ and $\beta \leqslant 1$, then, $\rho \to 1$ because $x_0 e^{\beta\mu} < 1$, so that $x_0 < e^{-\beta\mu} \to 0$. Moreover from (4.1), because $1/x_0 \to \infty$, $\Sigma[(e^{\beta\mu}x_0)^n/n^\beta] \to \infty$ so that $e^{\beta\mu}x_0 \to 1$. Then

$$\rho = \frac{\Sigma(e^{\beta\mu}x_0)^n (n/n^\beta)}{\Sigma(e^{\beta\mu}x_0)^n (1/n^\beta) + \Sigma(e^{\beta\mu}x_0)^n (n/n^\beta)} \to 1.$$

The curves $p = p(\mu,\beta)$ and $\rho = \rho(\mu,\beta)$ are shown in Figs. 5(a)–(c) ($\beta \leqslant 1$, $1 < \beta \leqslant 2$ and $\beta > 2$, respectively).

The curves $p = p(\rho)$ give the diagrams in Fig. 6(a)–(c).

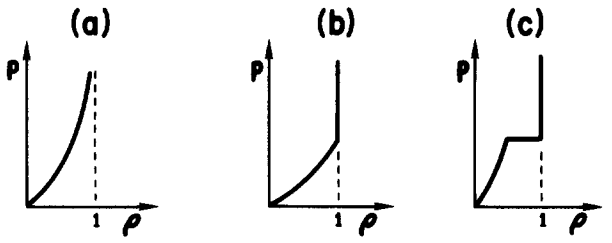

FIG. 6. Plot of p vs ρ for various ranges of β: (a) $\beta \leqslant 1$; (b) $1 < \beta \leqslant 2$; (c) $\beta > 2$.

V. MATHEMATICAL PRELIMINARIES IN THE CONDENSED PHASE

Our goal in the next several sections is to compute expected cluster numbers in the various phases. In this and the next section we lay the groundwork for the later calculations.

(a) Nonthermodynamic quantities. To compute *nonthermodynamic* quantities, we need to express Z_L and its asymptotics in a precise way. To do this, we have by (2.3)

$$\sum_{L>0} z^L Z_L = \frac{1}{1-z}\frac{1}{\Phi_{\beta,\mu}(z)}, \tag{5.1}$$

where

$$\Phi_{\beta,\mu}(z) \equiv 1 - z\phi_{\beta,\mu}(z).$$

First we consider the case $\mu > \mu_c(\beta)$. We can compute

$$Z_L = \frac{1}{2i\pi}\int_{\text{Circle of convergence}} \frac{(\Sigma_{L>0} z^L Z_L)}{z^{L+1}}\, dz.$$

But in this domain the circle of convergence is $z = e^{i\theta}e^{-\beta\mu}$, so that

$$Z_L = \frac{e^{\beta\mu L}}{2\pi}\int_0^{2\pi} \frac{1}{1-e^{i\theta}e^{-\beta\mu}}\frac{1}{\Phi_{\beta,\mu}(e^{-\beta\mu}e^{i\theta})} e^{-iL\theta}\, d\theta. \tag{5.2}$$

On this circle of convergence, the singularity of the function that we integrate is not a pole but the singularity of $1/\Phi_{\beta,\mu}$ which is the same as that of $\phi_{\beta,\mu}$.

(b) Singularity of $\phi_{\beta,\mu}$ $(1 < \beta < 2)$. We start with

$$\int_0^\infty e^{-kt} t^{\beta-1}\, dt = \frac{\Gamma(\beta)}{k^\beta},$$

so that

$$\phi_{\beta,\mu}(z) = \frac{1}{\Gamma(\beta)}\int_0^\infty \frac{e^{-t}e^{\beta\mu}z}{1-e^{-t}e^{\beta\mu}z} t^{\beta-1}\, dt. \tag{5.3}$$

We study the singularity in θ of this function for $z = e^{-\beta\mu}e^{i\theta}$. Substituting this value of z in (5.3) it is clear that the only singularity occurs at $\theta = 0$ (recall that $\beta > 1$). Define

$$1 + x = e^{i\theta} \quad (x \text{ small}), \tag{5.4}$$

$$\psi_\beta(x) = \int_0^\infty \frac{e^{-t}t^{\beta-1}\, dt}{1-e^{-t}-e^{-t}x} + x\int_0^\infty \frac{e^{-t}t^{\beta-1}}{1-e^{-t}-xe^{-t}}\, dt. \tag{5.5}$$

Then we see that only the first term can fail to be smooth for $x = 0$. Define

$$\tilde{\psi}_\beta(x) = \int_0^\infty e^{-t}\frac{t^{\beta-1}\, dt}{1-e^{-t}-e^{-t}x}. \tag{5.6}$$

Let $x = e^{i\theta} - 1 \sim ix'$ with x' real for θ small. Then

$$\begin{aligned}
\tilde{\psi}_\beta(x) - \tilde{\psi}_\beta(0) &= \int_0^\infty e^{-t}t^{\beta-1}\left(\frac{1}{1-e^{-t}ix'e^{-t}} - \frac{1}{1-e^{-t}}\right) dt \\
&= ix'\int_0^\infty \frac{t^{\beta-1}}{(e^t-1)(e^t-1-ix')}\, dt \\
&= ix'\int_0^\infty \frac{t^{\beta-1}(e^t-1+ix')}{(e^t-1)((e^t-1)^2+x'^2)}\, dt \\
&= ix'\int_0^\infty \frac{t^{\beta-1}}{(e^t-1)^2+x'^2}\, dt - x'^2\int_0^\infty \frac{t^{\beta-1}\, dt}{(e^t-1)((e^t-1)^2+x'^2)}.
\end{aligned}$$

We continue with a change of the variable $t = x'u$ $(x' > 0)$

$$= ix'^{\beta-1}\int_0^\infty \frac{u^{\beta-1}\, du}{[(e^{x'u}-1)/x']^2+1} - x'^{\beta-1} \times \int_0^\infty \frac{u^{\beta-1}\, du}{[(e^{x'u}-1)/x'][[(e^{x'u}-1)/x']^2+1]}.$$

If $x' \to 0$, $(e^{x'u}-1)/x' \downarrow u$, and thus for $x' > 0$:

$$\tilde{\psi}_\beta(x) - \tilde{\psi}_\beta(0) \sim x'^{\beta-1}\left\{i\int_0^\infty \frac{u^{\beta-1}\, du}{u^2+1} - \int_0^\infty \frac{u^{\beta-1}\, du}{u(u^2+1)}\right\} \quad (\text{here } \beta > 1). \tag{5.7}$$

For $x' < 0$, we define $x' = -|x'|$, $t = |x'|u$ and find

$$\tilde{\psi}_\beta(x) - \tilde{\psi}_\beta(0) \sim |x'|^{\beta-1}\left\{ -i\int_0^\infty \frac{u^{\beta-1}\,du}{u^2+1} - \int_0^\infty \frac{u^{\beta-1}\,du}{u(u^2+1)}\right\}. \quad (5.8)$$

Define

$$C_\beta = i\int_0^\infty \frac{u^{\beta-1}\,du}{u^2+1} - \int_0^\infty \frac{u^{\beta-1}\,du}{u(u^2+1)}$$

and assume $1<\beta\leqslant 2$. Then from (5.3) to (5.6)

$$\begin{aligned}
&\phi_{\beta,\mu}(e^{-\beta\mu}e^{i\theta}) = [1/\Gamma(\beta)]\tilde{\psi}_\beta(x) + O(x),\\
&x'\sim\theta \quad (e^{i\theta}-1\sim ix'),\\
&\phi_{\beta,\mu}(e^{-\beta\mu}e^{i\theta}) - \phi_{\beta,\mu}(e^{-\beta\mu})\\
&\quad\sim\begin{cases}[1/\Gamma(\beta)]C_\beta\theta^{\beta-1} & (\theta>0),\\ [1/\Gamma(\beta)]\overline{C}_\beta|\theta|^{\beta-1} & (\theta<0).\end{cases}
\end{aligned} \quad (5.9)$$

(c) Fourier coefficient of $\phi_{\beta,\mu}$, $1<\beta<2$. Let us start with a C^∞ function f of $\theta\in[-\pi,+\pi]$ except at 0 where it is only Hölder continuous with

$$f(\theta)\sim\begin{cases}A\theta^{\beta-1}, & \theta>0,\\ B|\theta|^{\beta-1}, & \theta<0.\end{cases}$$

Its Fourier coefficient

$$\hat{f}(n) = \int_{-\pi}^{+\pi} e^{-in\theta}f(\theta)\,\frac{d\theta}{2\pi}$$

has its growth controlled only by the behavior of $f(\theta)$ for $\theta\sim 0$. From Erdelyi[5] (p. 48) we deduce

$$\hat{f}(n)\sim[\Gamma(\beta)/2\pi n^\beta]\{-Ae^{-i\pi(\beta-2)/2} + Be^{i\pi\beta/2}\} \quad (5.10)$$

From (5.2), it follows that we must check the behavior at $\theta=0$ of the integrand

$$\begin{aligned}
&\frac{1}{1-e^{i\theta}e^{-\beta\mu}}\,\frac{1}{\Phi_{\beta,\mu}(e^{-\beta\mu}e^{i\theta})} - \frac{1}{1-e^{-\beta\mu}}\,\frac{1}{\Phi_{\beta,\mu}(e^{-\beta\mu})}\\
&\quad=\frac{1}{1-e^{-\beta\mu}}\left\{\frac{1}{1-e^{-\beta\mu}e^{i\theta}\phi_{\beta,\mu}(e^{-\beta\mu}e^{i\theta})} - \frac{1}{1-e^{-\beta\mu}\phi_{\beta,\mu}(e^{-\beta\mu})}\right\} + O(\theta)\\
&\quad\sim\frac{1}{1-e^{-\beta\mu}}\,\frac{e^{-\beta\mu}[\phi_{\beta,\mu}(e^{-\beta\mu}e^{i\theta}) - \phi_{\beta,\mu}(e^{-\beta\mu})]}{[1-e^{-\beta\mu}\phi_{\beta,\mu}(e^{-\beta\mu})]^2} + O(\theta)\\
&\quad\sim\frac{e^{-\beta\mu}}{(e^{\beta\mu}-1)[1-e^{-\beta\mu}\phi_{\beta,\mu}(e^{-\beta\mu})]^2}\times\begin{cases}A_\beta\theta^{\beta-1}, & \theta>0,\\ B_\beta|\theta|^{\beta-1}, & \theta<0,\end{cases}
\end{aligned}$$

where A_β and B_β are universal functions of β, namely, basically the Hölder modulus of continuity of $\phi_{\beta,\mu}(e^{-\beta\mu}e^{i\theta})$ for $\theta\sim 0^+$ or $\theta\sim 0^-$. Then using (5.2) and (5.10) we see that for $L\to\infty$

$$Z_L\sim\frac{e^{\beta\mu L}}{L^\beta}\,\frac{e^{-\beta\mu}}{(1-e^{-\beta\mu})[1-e^{-\beta\mu}\phi_{\beta,\mu}(e^{-\beta\mu})]^2}K_\beta, \quad (5.11)$$

where K_β is a universal function of β depending only on the Hölder modulus of continuity of $\phi_{\beta,\mu}(e^{-\beta\mu}(e^{i\theta})$ for small θ.

(d) The case $2<\beta<3$. In this case, $\Phi_{\beta,\mu}(e^{-\beta\mu}e^{i\theta})$ has a derivative everywhere but only the first derivative has a singularity at $\theta=0$. More precisely

$$\begin{aligned}
&\frac{d}{d\theta}(\Phi_{\beta,\mu}^{-1}(e^{-\beta\mu}e^{i\theta}))\\
&\quad=-\frac{1}{(\Phi_{\beta,\mu}(e^{-\beta\mu}e^{i\theta}))^2}\,\frac{d}{d\theta}\Phi_{\beta,\mu}(e^{-\beta\mu}e^{i\theta})
\end{aligned} \quad (5.12)$$

and

$$\begin{aligned}
\frac{d}{d\theta}\Phi_{\beta,\mu}(e^{-\beta\mu}e^{i\theta}) &= \frac{d}{d\theta}\left\{1-e^{-\beta\mu}e^{i\theta}\sum_{n\geq 1}\frac{e^{in\theta}}{n^\beta}\right\}\\
&= -e^{-\beta\mu}ie^{i\theta}\phi_{\beta,\mu}(e^{-\beta\mu}e^{i\theta})\\
&\quad - e^{-\beta\mu}e^{i\theta}i\phi_{\beta-1,\mu}(e^{-\beta\mu}e^{i\theta}),
\end{aligned} \quad (5.13)$$

so that $\Phi_{\beta,\mu}$ is C^∞ everywhere except at 0 and at 0 it is of class $C^{\beta-2}$. (It has a first derivative at 0 which is Hölder continuous of order $\beta-2$.)

To treat (5.2) we write $e^{iL\theta} = (1/Li)(d/d\theta)(e^{-iL\theta})$ and integrate by parts

$$Z_L = \frac{e^{\beta\mu L}}{Li}\int_0^{2\pi}\left[x\,\frac{d}{d\theta}\left(\frac{1}{1-e^{i\theta}e^{-\beta\mu}}\,\frac{1}{\Phi_{\beta,\mu}(e^{-\beta\mu}e^{i\theta})}\right)\right] \times e^{-iL\theta}\,d\theta.$$

Then the behavior for large L is controlled by the Hölder continuity at 0 of this derivative. Looking at (5.12) and (5.13) we see that the singular part of this derivative is

$$\frac{e^{-\beta\mu}}{(\Phi_{\beta,\mu}(e^{-\beta\mu}))^2}\,\phi_{\beta-1,\mu}(e^{-\beta\mu}e^{i\theta})\,\frac{1}{1-e^{-\beta\mu}}.$$

But this function is Hölder continuous at $\theta=0$ for order $\beta-2$, and thus

$$Z_L\sim\frac{e^{\beta\mu L}}{L^\beta}\,\frac{e^{-\beta\mu}}{(1-e^{-\beta\mu})(\Phi_{\beta,\mu}(e^{-\beta\mu}))^2}K_{\beta-1}, \quad (5.14)$$

where $K_{\beta-1}$ depends only on the modulus of continuity at 0 of $\phi_{\beta-1,\mu}(e^{-\beta\mu})$.

VI. MATHEMATICAL PRELIMINARIES IN THE NONCONDENSED PHASE

(a) In the noncondensed phase ($\beta<1$ or $\mu<\mu_c$) the function $\Sigma_{L>0} z^L Z_L$ has a simple pole on the circle of convergence because this is $1/(1-z)$ $[1/\Phi_{\beta,\mu}(z)]$ and $\Phi_{\beta,\mu}(z)$ goes linearly to 0 for $z\to x_0$ where x_0 is the root of

$$1/x_0 = \phi_{\beta,\mu}(x_0).$$

In this case we cannot apply (5.2) directly. Let us consider

$$\sum_{L>0} \frac{x^{L+1}}{L+1} Z_L = \int_0^x \frac{dx'}{(1-x')\Phi_{\beta,\mu}(x')} \equiv G_\beta(x), \quad (6.1)$$

which has logarithmic behavior for $x = x_0$, namely,

$$G_\beta(x) \sim \frac{1}{1-x_0} \frac{1}{\Phi'_{\beta,\mu}(x_0)} \log(x-x_0). \quad (6.2)$$

Then

$$\frac{Z_L}{L+1} = \frac{x_0^{-(L+1)}}{2\pi} \int_0^{2\pi} G_\beta(x_0 e^{i\theta}) e^{i(L+1)\theta}\, d\theta. \quad (6.3)$$

(b) One must be careful about the determination of $\log(x-x_0)$ for $x = x_0 e^{i\theta}$. The correct branch is

$$\log(x_0 e^{i\theta} - x_0) \sim \log|\theta| + i(\operatorname{sgn}\theta)(\pi/2) + \text{regular term}$$

and thus if $h(\theta)$ is a C^∞ function of θ which is 1 at $\theta = 0$:

$$\int_{-\pi}^{+\pi} e^{-in\theta} \left(\log|\theta| + i(\operatorname{sgn}\theta)\frac{\pi}{2}\right) h(\theta) d\theta \sim \frac{2\pi}{n},$$

for large n. (This is proved in the Appendix.) In particular coming back to (6.3), we obtain

$$Z_L \sim x_0^{-(L+1)} / [(1-x_0)\Phi'_{\beta,\mu}(x_0)] \quad (6.4)$$

[x_0 is the root of $1/x_0 = \phi_{\beta,\mu}(x_0)$].

VII. COMPUTATION OF THE NUMBERS OF CLUSTERS

A. Notation

We denote by $\langle\cdots\rangle_L$ the thermodynamic average of the quantity $\cdots$ and by $[\cdots]_L = Z_L\langle\cdots\rangle_L$ the non-normalized average of this quantity. It is usually simpler to compute $[\cdots]_L$ and to obtain $\langle\cdots\rangle_L$ as $[\cdots]_L/Z_L$.

B. Computation of $[N_n]_L$ and $[P]_L$: Generating functions

Introduce for each n a chemical potential μ_n so that $E(n) = \log n - \mu_n n$ and thus

$$[N_n]_L = \sum_{\{N_n,P\}}{}' N_n \frac{(\Sigma N_n)!}{\Pi N_n} \prod_n e^{-\beta E(n) N_n} e^{-\beta\nu P}\Big|_{\mu_n=\mu}$$
$$= +\frac{1}{n\beta}\frac{\partial Z_L}{\partial \mu_n}\Big|_{\mu_n=\mu},$$

with the prime as in Eq. (2.2). Using Eq. (2.3) we therefore have

$$\sum_{L>0} x^L [N_n]_L = \frac{-1}{n\beta}\frac{\partial}{\partial\mu_n}\left(\sum x^L Z_L\right)\Big|_{\mu_n=\mu}$$
$$= \frac{1}{n\beta}\frac{1}{1-x}\frac{1}{\Phi_{\beta,\mu}(x)^2} x^{n+1} e^{-\beta E(n)}\beta n$$

or

$$\sum_{L>0} x^L [N_n]_L = \frac{1}{1-x}\frac{1}{\Phi_{\beta,\mu}(x)^2} x^{n+1} e^{-\beta E(n)}. \quad (7.1)$$

Also

$$[P]_L = \frac{-1}{\beta}\frac{\partial}{\partial\nu}(Z_L)\Big|_{\nu=0}$$

and using (2.3)

$$\sum_{L>0} x^L [P]_L = \frac{1}{\Phi_{\beta,\mu}(x)}\frac{x}{(1-x)^2} = \frac{x}{1-x}\sum_{L>0} x^L Z_L. \quad (7.2)$$

C. First case: Noncondensed phase ($\beta<1$ or $\mu<\mu_c$)

In this case we evaluate $[N_n]_L$ by integrating on the circle of convergence of radius x_0. But (7.1) has a double pole on the circle of convergence at $z = x_0$ so that we must integrate twice

$$\sum_{L>0} \frac{x^{L+2}}{(L+1)(L+2)} [N_n]_L$$
$$= \left(\int_0^x dx' \int_0^{x'} dx'' \frac{1}{1-x''}\frac{x''^{n+1}}{\Phi_{\beta,\mu}(x'')^2}\right) e^{-\beta E(n)}.$$

But $\Phi_{\beta,\mu}(x'') \sim \Phi'_{\beta,\mu}(x_0)(x''-x_0)$, and thus at x_0 the singularity of the function on the right-hand side is

$$-e^{-\beta E(n)}[x_0^{n+1}/(1-x_0)]$$
$$\times(\Phi'_{\beta,\mu}(x_0))^{-2}\log(x''-x_0)$$

around x_0 and

$$\frac{[N_n]_L}{(L+1)(L+2)}$$
$$= -\frac{x_0^{-(L+2)}}{2\pi}\int_0^{2\pi} e^{-i(L+2)\theta}$$
$$\times\left(\sum_{L>0}\frac{(x_0 e^{i\theta})^{L+2}}{(L+1)(L+2)}[N_n]_L\right)d\theta$$
$$\sim -\frac{x_0^{-(L+2)}}{L+2} e^{-\beta E(n)} \frac{x_0^{n+1}}{(1-x_0)\Phi'_{\beta,\mu}(x_0)^2}$$

for L large and using (6.4)

$$\langle N_n\rangle_L = \frac{[N_n]_L}{Z_L} \sim -Lx_0^{-1}\frac{e^{-\beta E(n)} x_0^{n+1}}{\Phi'_{\beta,\mu}(x_0)}.$$

The number of clusters per unit volume is thus

$$\frac{\langle N_n\rangle_L}{L} \sim -\frac{e^{-\beta E(n)} x_0^n}{\Phi'_{\beta,\mu}(x_0)}. \quad (7.3)$$

Remark 1. A check: The quantity $\Sigma n\langle N_n\rangle_L/L$ should be the density and by Eq. (7.2) it is equal to

$$-\frac{1}{\Phi'_{\beta,\mu}(x_0)}\sum n\frac{x_0^n e^{\beta\mu n}}{n^\beta}.$$

But

$$-\Phi'_{\beta,\mu}(x_0) = +\phi_{\beta,\mu}(x_0) + \sum n\frac{x_0^n e^{\beta\mu n}}{n^\beta}$$
$$= \frac{1}{x_0} + \sum n x_0^n \frac{e^{\beta\mu n}}{n^\beta}$$

[because $1/x_0 = \phi_{\beta,\mu}(x_0)$ by definition of x_0] so that

$$\sum n \frac{\langle N_n \rangle_L}{L} = \frac{\Sigma n x_0^n (e^{\beta\mu n}/n^\beta)}{1/x_0 + \Sigma n x_0^n (e^{\beta\mu n}/n^\beta)}$$

which checks correctly with (4.2).

Remark 2. We see also from (7.2) that the ratio of cluster size expectations is given by

$$\frac{\langle N_n \rangle_L}{\langle N_m \rangle_L} \sim \frac{e^{-\beta E(n)} x_0^n}{e^{-\beta E(m)} x_0^m} = \frac{(1/n^\beta) e^{(\beta\mu + \log x_0)n}}{(1/m^\beta) e^{(\beta\mu + \log x_0)m}}. \quad (7.4)$$

This means that from the standpoint of statistics the multi-urn problem can be thought of as an effective single-urn model at a temperature β^{-1} with an effective chemical potential

$$\mu_{\text{eff}} = \mu + \log x_0/\beta$$

(recall that $x_0 < 1$, so $\log x_0 < 0$). Moreover, since $x_0 < e^{-\beta\mu}$ (see Fig. 1) it follows that $\mu_{\text{eff}} < 0$. Returning to (7.2) we also see that

$$\langle P \rangle_L \sim x_0/(1 - x_0), \quad (7.5)$$

which means that at thermal equilibrium there is only a finite number of particles in the reservior (an *absolute* finite number when $L \to \infty$, that is, this number does not increase with increasing L).

D. Second case: Condensed phase ($1 < \beta < 2$ and $\mu > \mu_c$)

In this case the radii of circles of convergence for the series (7.1) and (7.2) are given by $e^{-\beta\mu}$. From (7.1) we obtain

$$[N_n]_L = \frac{e^{\beta\mu L}}{2\pi} \left(\int_0^{2\pi} \frac{e^{-\beta\mu(n+1)} e^{i(n+1)\theta} e^{-iL\theta}}{(1 - e^{-\beta\mu} e^{i\theta}) \Phi_{\beta,\mu}(e^{-\beta\mu} e^{i\theta})^2} d\theta \right) e^{-\beta E(n)}$$

We have to obtain the modulus of continuity at $\theta = 0$ of the function in the integral; this will be

$$\begin{aligned} &\frac{e^{-\beta\mu(n+1)}}{1 - e^{-\beta\mu} e^{i\theta}} \left(\frac{1}{\Phi_{\beta,\mu}(e^{-\beta\mu} e^{i\theta})^2} - \frac{1}{\Phi_{\beta,\mu}(e^{-\beta\mu})^2} \right) \\ &\quad \sim \frac{e^{-\beta\mu(n+1)}}{1 - e^{-\beta\mu} e^{i\theta}} \frac{1}{\Phi_{\beta,\mu}(e^{-\beta\mu})^4} \\ &\qquad \times [\Phi_{\beta,\mu}(e^{-\beta\mu}) - \Phi_{\beta,\mu}(e^{-\beta\mu} e^{i\theta})] \\ &\qquad \times [\Phi_{\beta,\mu}(e^{-\beta\mu}) + \Phi_{\beta,\mu}(e^{-\beta\mu} e^{i\theta})] \\ &\quad \sim \frac{e^{-\beta\mu(n+1)}}{1 - e^{-\beta\mu}} \frac{2}{\Phi_{\beta,\mu}(e^{-\beta\mu})^3} e^{-\beta\mu} \\ &\qquad \times [\phi_{\beta,\mu}(e^{-\beta\mu} e^{i\theta}) - \phi_{\beta,\mu}(e^{-\beta\mu})], \end{aligned}$$

so that

$$\begin{aligned} [N_n]_L \sim &\frac{e^{\beta\mu L}}{(L-n)^\beta} \\ &\times \frac{e^{-\beta\mu(n+1)} 2 e^{-\beta\mu}}{(1 - e^{-\beta\mu}) \Phi_{\beta,\mu}(e^{-\beta\mu})^3} K_\beta e^{-\beta E(n)}, \end{aligned}$$

with the same K_β as in (5.11) so that

$$\langle N_n \rangle_L \sim \frac{2 e^{-\beta\mu(n+1)}}{\Phi_{\beta,\mu}(e^{-\beta\mu})} e^{-\beta E(n)}$$

for fixed n which is also

$$\langle N_n \rangle_L \sim 2 e^{-\beta\mu} \frac{1}{n^\beta} \frac{1}{\Phi_{\beta,\mu}(e^{-\beta\mu})}. \quad (7.6)$$

This is an absolute number when $L \to \infty$. We also need $[P]_L$ and $\langle P \rangle_L$ from (7.2); we obtain immediately

$$\langle P \rangle_L \sim e^{-\beta\mu}/(1 - e^{-\beta\mu}) \quad (7.7)$$

E. Third case: Condensed phase $2 < \beta < 3$ and $\mu > \mu_c$

We need again the singularity of the right member of (7.1) and use an integration by parts; the main singularity will come from the derivative of $1/\Phi_{\beta,\mu}^2$ and will give

$$\frac{2 e^{-\beta\mu}}{\Phi_{\beta,\mu}(e^{-\beta\mu})^3} \phi_{\beta-1,\mu}(e^{-\beta\mu} e^{i\theta}) \frac{e^{-\beta\mu(n+1)}}{1 - e^{-\beta\mu}}$$

and finally this will give us the same answer for $\langle N_n \rangle_L$ as in the case $1 < \beta < 2$:

$$\langle N_n \rangle_L \sim 2 e^{-\beta\mu} (1/n^\beta) [1/\Phi_{\beta,\mu}(e^{-\beta\mu})] \quad (7.8)$$

and also

$$\langle P \rangle_L \sim e^{-\beta\mu}/(1 - e^{-\beta\mu}). \quad (7.9)$$

VIII. TOTAL NUMBER OF CLUSTERS IN THE CONDENSED PHASE

A. Let us define

$\pi_n(L) =$ *unnormalized probability* that $\Sigma_{k>1} N_k = n$.

To compute this quantity, we remark that $\pi_p(L)$ is the constrained partition function

$$\pi_p(L) = \sum_{\substack{\{N_n, P\} \\ \Sigma N_n = p}} \frac{(\Sigma_n N_n)!}{\Pi_n N_n!} \prod_n e^{-\beta E(n) N_n} e^{-\beta\nu P}$$

so that necessarily $L \geqslant p$ and

$$\begin{aligned} &\sum_{L \geqslant p} x^L \pi_p(L) \\ &\quad = \sum_{L \geqslant p} x^p \sum_{\substack{\{N_n, P\} \\ \Sigma N_n = p}} \frac{p!}{\Pi N_n!} \prod_n (e^{-\beta E(n)} x^n)^{N_n} (e^{-\beta\nu} x)^P \\ &\quad = x^p \sum_{\{N_n, P\}} \frac{p!}{\Pi N_n!} \prod_n (e^{-\beta E(n)} x^n)^{N_n} (e^{-\beta\nu} x)^P \\ &\quad = \frac{x^p}{1 - e^{-\beta\nu} x} \left(\sum_n e^{-\beta E(n)} x^n \right)^p. \end{aligned}$$

Summarizing

$$\sum_{L \geqslant p} x^L \pi_p(L) = \frac{[x \phi_{\beta,\mu}(x)]^p}{1 - e^{-\beta\nu} x}. \quad (8.1)$$

From this, we deduce as usual, in the condensed phase

$$\pi_p(L) = \frac{e^{\beta\mu L}}{2\pi} \int_0^{2\pi} \frac{(e^{-\beta\mu} e^{i\theta})^p [\phi_{\beta,\mu}(e^{-\beta\mu} e^{i\theta})]^p e^{-iL\theta}}{1 - e^{-\beta\mu} e^{i\theta}} d\theta.$$

Again, we suppose $1 < \beta < 2$. Then the singularity of the function that we integrate is the singularity of $\phi_{\beta,\mu}(e^{-\beta\mu} e^{i\theta})$ at $\theta = 0$ and this is controlled by

$$[e^{-p\beta\mu}/(1 - e^{-\beta\mu})] [\phi_{\beta,\mu}(e^{-\beta\mu} e^{i\theta})^p - \phi_{\beta,\mu}(e^{-\beta\mu})^p],$$

which is controlled finally by

$$[e^{-p\beta\mu}/(1-e^{-\beta\mu})]p(\phi_{\beta,\mu}(e^{-\beta\mu}))^{p-1}$$
$$\times|\phi_{\beta,\mu}(e^{-\beta\mu}e^{i\theta})-\phi_{\beta,\mu}(e^{-\beta\mu})|.$$

Therefore, for $L\to\infty$

$$\pi_p(L)\sim\frac{e^{\beta\mu(L-p)}}{(L-p)^{\beta}}\frac{1}{1-e^{\beta\mu}}p(\phi_{\beta,\mu}(e^{-\beta\mu}))^{p-1}K_{\beta},$$

where the K_β is the same as in (5.11). Then the normalized probability is

$$\pi_p(L)/Z_L\sim e^{-\beta\mu(p-1)}p(\phi_{\beta,\mu}(e^{-\beta\mu}))^{p-1}$$
$$\times(\Phi_{\beta,\mu}(e^{-\beta\mu}))^2. \tag{8.2}$$

Now we compute

$$\sum_p\lim_{L\to\infty}\frac{\pi_p(L)}{Z_L}$$
$$=\left[\sum_p p[e^{-\beta\mu}\phi_{\beta,\mu}(e^{\beta\mu})]^{p-1}\right]\Phi_{\beta,\mu}(e^{-\beta\mu})^2.$$

But

$$\sum p\xi^{p-1}=\frac{d}{d\xi}\left(\frac{1}{1-\xi}\right)=\frac{1}{(1-\xi)^2}$$

so that by definition of $\Phi_{\beta,\mu}(e^{-\beta\mu})$, we obtain 1.

B. *The average number of clusters is*

$$\sum_{p>0}p\frac{\pi_p(L)}{Z_L}$$
$$\sim\left[\sum_p p^2[\phi_{\beta,\mu}(e^{-\beta\mu})e^{-\beta\mu}]^{p-1}\right]\Phi_{\beta,\mu}(e^{-\beta\mu})^2.$$

But

$$\sum p^2\xi^{p-1}=\sum p(p-1)\xi^{p-1}+\sum p\xi^{p-1}$$
$$=\frac{2\xi}{(1-\xi)^3}+\frac{1}{(1-\xi)^2}=\frac{1+\xi}{(1-\xi)^3},$$

so that by definition of $\xi=e^{-\beta\mu}\phi_{\beta,\mu}(e^{-\beta\mu})$ and of $\Phi_{\beta,\mu}(e^{-\beta\mu})$

$$\sum_{p>0}p\frac{\pi_p(L)}{Z_L}\sim\frac{1+e^{-\beta\mu}\phi_{\beta,\mu}(e^{-\beta\mu})}{\Phi_{\beta,\mu}(e^{-\beta\mu})}\ (>1). \tag{8.3}$$

C. Now let us look at

$$\sum_n\langle N_n\rangle_L,$$

which is the average number of clusters of finite size. By (7.8) this is

$$\sum\langle N_n\rangle_L\sim\frac{1}{\Phi_{\beta,\mu}(e^{-\beta\mu})}2\left(\sum_n\frac{1}{n^\beta}\right)e^{-\beta\mu}$$
$$=\frac{1}{\Phi_{\beta,\mu}(e^{-\beta\mu})}2e^{-\beta\mu}\phi_{\beta,\mu}(e^{-\beta\mu}). \tag{8.4}$$

Then we see that

$$\sum_{p>0}p\frac{\pi_p(L)}{Z_L}-\sum_n\langle N_n\rangle_L=\frac{1-e^{-\beta\mu}\phi_{\beta,\mu}(e^{-\beta\mu})}{\Phi_{\beta,\mu}(e^{-\beta\mu})}=1, \tag{8.5}$$

which means that on the average there is one cluster of infinite size and a certain total number (finite and independent of L) of clusters of finite size.

Remark: Looking at the expression (8.4) for the average number of clusters of finite size, we see that this number tends to 0 like $e^{-\beta\mu}$ if $\mu\to+\infty$.

IX. THE AVERAGE NUMBER OF PARTICLES PER CLUSTER

There are ΣnN_n particles in a configuration, and also ΣN_n clusters. The average cluster of the configuration has a size l

$$l\equiv\sum_n nN_n\left(\sum_n N_n\right)^{-1}$$

(average size of a cluster of a given configuration) (9.1)

and the unnormalized thermal average is thus

$$[l]_L\equiv\left[\frac{\Sigma nN_n}{\Sigma N_n}\right]_L$$
$$=\sum_{\substack{\{N_n\}\\ \Sigma nN_n+\Sigma N_{n<L}}}\frac{(\Sigma N_k)!}{\Pi N_k!}\frac{\Sigma nN_n}{\Sigma N_n}\prod_k e^{-\beta N_kE(k)}e^{-\beta\nu P}.$$

We can now write a generating function (which is not the same as the true cluster size generating function)

$$\sum_{L>0}x^L[l]_L=\sum_n\sum_{\substack{N_k>0\\ N_n>1\\ P>0}}n\frac{(\Sigma N_k-1)!}{\Pi_{k\neq n}N_k!(N_n-1)!}$$
$$\times\prod_k(e^{-\beta E(k)}x^{k+1})^{N_k}(e^{-\beta\nu}x)^P.$$

As usual we set $\nu=0$, and perform the sum over $P\geqslant 0$ which gives $1/(1-x)$. Then we sum over n first to get by slight rearrangement

$$\sum_{L>0}x^L[l]_L$$
$$=\sum_n\frac{ne^{-\beta E(n)}x^{n+1}}{1-x}\sum_{\substack{N_k>0\\ N_n>1}}\frac{(\Sigma N_k-1)!}{\Pi_{k\neq n}N_k!(N_n-1)!}$$
$$\times\prod_{k\neq n}(e^{-\beta E(k)}x^{k+1})^{N_k}(e^{-\beta E(n)}x^{n+1})^{N_n-1}.$$

But then we see that multiplying the term $ne^{-\beta E(n)}x^{n+1}$ there appears exactly the quantity

$$\sum_{N_k>0}\frac{(\Sigma N_k)!}{\Pi N_k!}\prod_k(e^{-\beta E(k)}x^{k+1})^{N_k}=\frac{1}{1-x\phi_{\beta,\mu}(x)},$$

so that

$$\sum_{L>0}x^L[l]_L=\frac{1}{(1-x)(1-x\phi_{\beta,\mu}(x))}$$
$$\times\sum_n ne^{-\beta E(n)}x^{n+1}$$
$$=\frac{1}{(1-x)(1-x\phi_{\beta,\mu}(x))}x^2\frac{d}{dx}\phi_{\beta,\mu}(x).$$

Now, if we assume that we are in the noncondensed phase [so that $\beta < 1$ or $\beta \geqslant 1$ and $\mu < \mu_c(\beta)$], we can immediately adapt the computation of Sec. IV to obtain the normalized thermal average of l:

$$\langle l \rangle_L \sim x_0^2 \frac{d}{dx} \phi_{\beta\mu} \mid_{x=x_0} = \sum_{n>1} n e^{-\beta E(n)} x_0^{n+1}. \quad (9.2)$$

On the other hand, we can compute

$$\sum n \langle N_n \rangle_L \left(\sum \langle N_n \rangle_L \right)^{-1}.$$

But by (7.4)

$$\langle N_n \rangle_L / \langle N_m \rangle_L \sim e^{-\beta E(n)} x_0^n / e^{-\beta E(m)} x_0^m,$$

so that

$$\frac{\Sigma n \langle N_n \rangle_L}{\Sigma \langle N_m \rangle_L} \sim \frac{\Sigma n e^{-\beta E(n)} x_0^n}{\Sigma e^{-\beta E(n)} x_0^n}. \quad (9.3)$$

But $1/x_0 = \Sigma e^{-\beta E(n)} x_0^n$ (by the definition of x_0), so that

$$\frac{\Sigma n \langle N_n \rangle_L}{\Sigma \langle N_n \rangle_L} \sim \sum n e^{-\beta E(n)} x_0^{n+1}.$$

Comparing (9.2) and (9.3) we find

$$\langle l \rangle_L = \left\langle \frac{\Sigma n N_n}{\Sigma N_n} \right\rangle_L \sim \frac{\Sigma n \langle N_n \rangle_L}{\Sigma \langle N_n \rangle_L} \sim \sum_n n e^{-\beta E(n)} x_0^{n+1}. \quad (9.4)$$

This result is remarkable because it states that in the thermodynamic limit, if we compute the thermal average of the size of the average cluster ($\langle l \rangle_L$), then it is exactly the thermal average of the size of the effective urn with an effective chemical potential $\mu_{\text{eff}} = \mu + (1/\beta) \log x_0$. It is also remarkable that for large L

$$\left\langle \frac{\Sigma n N_n}{\Sigma N_n} \right\rangle_L = \frac{\Sigma n \langle N_n \rangle_L}{\Sigma \langle N_n \rangle_L}.$$

X. DISTRIBUTION OF PARTICLES IN THE RESERVOIR

In this section we compute the probability distribution of the number of particles in the reservoir. We define

$$R_k = \lim_{L \to \infty} \langle \delta(P - k) \rangle_L, \quad (10.1)$$

where P is the number of particles in the reservoir. Thus R_k is the probability that the reservoir contains k particles. It is clear that

$$R_k = \lim_{L \to \infty} Z_{L,k} / Z_L,$$

where Z_L is the usual partition function and $Z_{L,k}$ is the constrained partition function, summing only over configurations such that the reservoir contains exactly $P = k$ particles. It follows, as in Sec. I, that

$$\sum_{L>k} x^L Z_{L,k} = \frac{x^k}{1 - x\phi_{\beta,\mu}(x)}. \quad (10.2)$$

We thus see that the difference between the generating function of Z_L given by (2.3) and the generating function of $Z_{L,k}$ given by (10.2) is the presence of $(1-x)^{-1}$ in (2.3) and of x^k in (10.2). From this, and the fact that the singularities of the generating function of Z_L and $Z_{L,k}$ are the same, we can infer immediately that

$$R_k = e^{-\beta p k}(1 - e^{-\beta p}) \quad (10.3)$$

both in the gaseous and condensed phases (in the latter case, p is simply μ).

XI. THE MEANING OF THE EQUATION OF STATE IN THE GAS PHASE

We now want to give a self-consistent interpretation of the thermodynamic state. Let us recall that for $\mu < \mu_c(\beta)$, we have defined

$$\frac{1}{x_0} = \phi_{\beta,\mu}(x_0) \equiv \sum e^{-\beta \log n + \beta\mu n} x_0^n. \quad (11.1)$$

On the other hand, we know [see formula (3.2)]

$$-\log x_0 = \lim_{L \to \infty} (1/L) \log Z_L \quad (11.2)$$

so that $+\log x_0/\beta$ is *the free energy per unit length of our system of clusters.*

Call $f = \log x_0/\beta$. Equation (11.1) can be rewritten as follows:

$$1 = \frac{\Sigma_n e^{-\beta(\log n - (\mu + f)n)}}{e^{-\beta f}}. \quad (11.3)$$

Equation (11.3) is a normalization equation that can be interpreted as follows: The initial system of clusters is equivalent to a unique urn with chemical potential $\mu + f$, where f is chosen so that the partition function of this effective urn is precisely $e^{-\beta f}$. *We have a unique urn with chemical potential $\mu + f$ with f precisely the free energy of this unique urn*: This condition is self-consistent with respect to f. Moreover, this condition also describes the complicated model of clusters with f free energy per unit length.

In fact, f is $-p$ where p is the pressure. This is because we are working in the grand canonical ensemble for the complicated model of clusters. Then the energy of an effective urn of size n is exactly the energy $E(n) = \log n - \mu n$ of a cluster of size n minus n times the free energy f per unit length, or *the energy of an effective urn of size n is exactly the energy $E(n) = \log n - \mu n$ of a cluster of size n plus n times the pressure* (pressure that this urn would get from its environment if it were embedded in the complicated model of clusters).

This discussion points to the following fact: After all, we can forget completely the complicated model of clusters, start with a bare urn $E(n) = \log n - \mu n$ and change its chemical potential μ to $\mu + f$ so that f is precisely the free energy of this new urn. This is always possible if either $\beta < 1$ or $\beta > 1$ and $\mu < \mu_c(\beta)$. We also know from Sec. II and Figs. 1 and 2, that $x_0 e^{\mu\beta} < 1$ but $x_0 = e^{\beta f}$ so $f + \mu < 0$ and our new effective urn has its new chemical potential $\mu + f < 0$, so it is stable and its partition function is normalizable (which we already know because this partition function is $e^{-\beta f}$).

We now want to justify in physical terms the equation of state (11.3). First we consider a single urn: If it were completely isolated it would have the energy

$$E(n) = \log n - \mu n$$

if it had size n. But it is not isolated; in particular it lives with other competing urns. We call these other urns the *bath*. (This "bath" is actually a particle reservoir, but we use the term "bath" to avoid confusion with the distinguished reservoir cluster defined earlier whose occupation number is P.) The bath induces a pressure p on the urn, so that each particle inside the urn has the extra energy p. There is thus a new energy of the urn in presence of the bath

$$E'(n) = \log n - \mu n + pn.$$

The correct partition function for the urn in contact with the bath will then be $\Sigma_n e^{-\beta E'(n)}$.

On the other hand, we now have an observer who can act on the urn by adding or taking away a particle. But the observer does not know what microscopic mechanism within the urn gives rise to $\log n - \mu n$. All that he knows is that the exterior, or the bath, gives a pressure p for each particle in the urn. He can get this by macroscopic or thermodynamic experiments; he cannot really look inside the urn to get $\log n - \mu n$. So, from the point of view of the observer, considering the urn as a whole, the partition function should be $e^{+\beta p}$. (We take the plus sign because this is the energy that the observer has to give to a particle in the urn to let it go out to the bath.) So $e^{\beta p}$ represents the macroscopic point of view of the observer making experiments on the urn as a whole and $\Sigma e^{-\beta E'(n)}$ is the microscopic point of view of the urn in contact with its bath: *The two expressions have to be equal, so* $e^{\beta p} = \Sigma_n e^{-\beta E'(n)}$, *which is Eq. (11.3).* In effect we have renormalized our urn.

Remark: This kind of reasoning appears in Landau and Lifshitz[6] when they argue that a thermodynamic system (here the urn) cannot distinguish between a fluctuation due to the thermal reservoir and a fluctuation due to the action of an observer. (This consideration underlies fluctuation-dissipation theorem as presented in that reference.) Here we have applied the same kind of reasoning to express the partition function in two different ways and to recover Eq. (11.3).

Remark: Let us now look at the case $\beta > 1$, $\mu > \mu_c(\beta)$, which corresponds to the condensed phase in the model of clusters. Then Eq. (11.3) has no real solution f. In fact, we know from Sec. II that the pressure p (which is also $-f$) is $p = \mu$ but $f = -\mu$ is clearly not a solution of (11.3) except for the limiting case $\mu = \mu_c(\beta)$ [because $\mu_c(\beta)$] is

$$e^{\beta\mu_c(\beta)} = \sum \frac{1}{n^\beta}$$

[formula (3.3)]. This means that the preceding reasoning for the noncondensed phase cannot be applied to the condensed phase, because in the condensed phase the system of clusters is filled just with one urn of size $\sim L - \epsilon$ plus a few urns of microscopic size and thus, there is not really any effective urn per unit length, so that this mean field type of reasoning does not make sense at all.

XII. DISCUSSION AND SUMMARY

Our model is intermediate in its complexity between the simple, single urn or cluster model[3,4] and the Fisher–Felderhof droplet model.[1,2] We allow an arbitrary number of growing and shrinking clusters but they live under an overall constraint (analogous to that imposed in the spherical model) that effectively couples them. The constraint is Eq. (2.1), which states that the total of cluster occupations, spaces "between" them and reservoir occupation is fixed. The thermodynamics of this system is similar to that of the single urn with an effective chemical potential shifted by the pressure [which at that point equals $\mu_c(\beta)$ with $\mu_c(\beta) = \beta^{-1} \times \log \Sigma_{n=1}^{\infty} n^{-\beta}$]. The phase diagram is shown in Fig. 7.

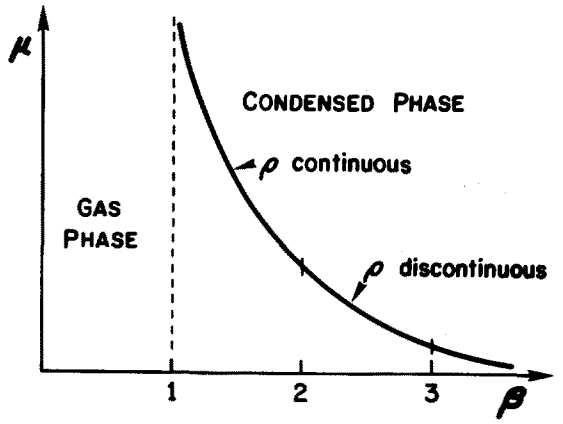

FIG. 7. Phase diagram in the β-μ plane. The phase transition curve is $\mu_c(\beta)$, given in Eq. (3.3). For $1 < \beta \leq 2$ the density ρ is continuous across the transition, but $\partial\rho/\partial\mu$ is not. For $2 < \beta$, ρ is discontinuous.

We then examine cluster size distribution in each of the phases. In the noncondensed phase the number of clusters of any given size is proportional to L, an appropriate thermodynamic scaling for a "gas" of clusters. In the condensed phase there are still some finite clusters, but only a limited number of them (that is, this number is finitely bounded as $L \to \infty$). All remaining balls or atoms are part of a single $O(L)$ cluster. In both cases, the reservoir is finitely (i.e., independent of L) occupied. Finally, we have given a mean field type argument to justify the equation of state.

As indicated in the Introduction, we intend to study a dynamical model whose equilibrium is that described in the present paper. As such, the properties we have here established will be features that must emerge in our dynamical model. Also, because of our interest in the metastable features of that dynamical model, the material developed here will be extended to examine the extent to which analytic continuation is a useful guide to metastable behavior.

ACKNOWLEDGMENTS

This work was supported in part by the U.S. National Science Foundation Grant Number 88-11106.

APPENDIX: EVALUATION OF THE INTEGRAL $\int_{-\pi}^{+\pi} e^{in\theta} \log|\theta| h(\theta)\, d\theta$ IN THE LIMIT OF LARGE n

If $\theta > 0$ and $h(\theta)$ is the C^∞ function, which is 1 at $\theta = 0$, we can say that

$$\int_0^{\pi} e^{ix\theta} \log \theta\, h(\theta) d\theta$$

$$\sim \int_0^{i\infty} \log u\, e^{ixu}\, du = \int_0^{\infty} \log(i\sigma) e^{-x\sigma} i\, d\sigma$$

$$= i \int_0^{\infty} \left(\log \sigma + i \frac{\pi}{2} \right) e^{-x\sigma}\, d\sigma$$

$$= -\frac{\pi}{2x} - i \frac{\log x}{x} + \frac{i}{x} \int_0^{\infty} \log u e^{-u}\, du,$$

$$\int_{-\pi}^{0} e^{in\theta} \log|\theta|\, h(\theta) d\theta$$

$$\sim -\int_0^{i\infty} \log(-\theta) e^{ix\theta}\, d\theta$$

$$\sim -i \int_0^{\infty} \log(-i\sigma) e^{-x\sigma}\, d\sigma$$

$$\sim -\frac{\pi}{2x} + i \frac{\log x}{x} - \frac{i}{x} \int_x^{\infty} \log u e^{-u}\, du,$$

so

$$\int_{-\pi}^{+\pi} e^{ix\theta} \log|\theta|\, h(\theta) d\theta \sim -\frac{\pi}{x}.$$

[1] M. E. Fisher and B. U. Felderhof, Ann. Phys. **58**, 176 (1970); **58**, 217 (1970); B. U. Felderhof and M. E. Fisher, *ibid.* **58**, 268 (1970).
[2] M. E. Fisher, Physics **3**, 255 (1967); B. U. Felderhof, Ann. Phys. **58**, 281 (1970).
[3] B. Gaveau and L. S. Schulman, J. Phys. A **20**, 2865 (1987).
[4] G. Roepstorff and L. S. Schulman, J. Stat. Phys. **34**, 35 (1984).
[5] A. Erdelyi, *Asymptotic Expansions* (Dover, New York, 1956).
[6] L. D. Landau and E. M. Lifshitz, *Statistical Physics* (Addison-Wesley, Reading, MA, 1958).

Solutions to higher Hamiltonians in the Toda hierarchies

L. A. Ferreira
Instituto de Fiisica Teórica-UNESP, Rua Pamplona 145, 01405, São Paulo, SP, Brazil

R. M. Londe
Centro Brasileiro de Pesquisas Fisicas-CBPF, Rua Dr. Xavier Sigaud 150, 22290, Rio de Janeiro, RJ, Brazil

(Received 31 January 1989; accepted for publication 5 July 1990)

A method is presented for constructing the general solution to higher Hamiltonians (nonquadratic in the momenta) of the Toda hierarchies of integrable models associated with a simple Lie group G. The method is representation independent and is based on a modified version of the Lax operator. It constitutes a generalization of the method used to construct the solutions of the Toda molecule models. The SL(3) and SL(4) cases are discussed in detail.

I. INTRODUCTION

Nonlinear systems in one and two space-time dimensions have been the object of a great interest in physics and mathematics. Their structures are, in general, very rich and very useful in the understanding of nonlinear properties of physical theories. The Toda molecule model[1] (TM) is one of these systems that has been studied quite extensively using several different approaches. The equations of motion for its one-dimensional version are given by

$$\frac{d^2\phi_a}{dt^2} = -\exp\left(\sum_b K_{ab}\phi_b\right), \quad a,b = 1,2,...,r. \tag{1.1}$$

For the case where the square, nonsingular matrix K is the Cartan matrix of a simple Lie group G, of rank r, these equations are completely integrable.[2,3] Their solutions have already been constructed.[4-6]

The Lie group G plays an important role in the study of the integrability properties of the TM equations. The solutions of Eqs. (1.1) can be viewed as some special geodesic motion on the symmetric space G^N/K, where G^N is the normal real form of G and K is its maximal compact subgroup.[5,7] When G^N/K is parametrized by the horospherical coordinates,[5,7] Eqs. (1.1) correspond, by a reduction procedure, to the radial part of the equations for the free motion on G^N/K. Although the TM equations are not invariant under continuous transformations associated with G, the invariance of the geodesic motion under left translations on G^N/K by G^N is, in some sense, hidden in (1.1).

Perhaps the richest structures of the TM equations associated with this hidden symmetry are contained in the so-called fundamental Poisson bracket relation (FPR) or the classical Yang–Baxter equation,[8,2,7,9,10] which relates two bracket structures. On one side is the Poisson bracket between the entries of a matrix operator A and on the other is the Lie bracket or commutator between A and a constant operator $\mathbb{P}$. In a tensor product notation, the FPR for the TM reads[2,7]

$$\{A \otimes A\}_{PB} = -[\mathbb{P}, A \otimes 1 + 1 \otimes A]. \tag{1.2}$$

The operator A lives in the subspace of the Lie algebra of G^N, which is the tangent space to G^N/K at unity, and it is given by

$$A = \frac{1}{2}\sum_{a,b=1}^{r} g_{ab}^{-1} p_a H_b + \frac{1}{2}\sum_{a=1}^{r} \exp\left[\frac{1}{2}\sum_{b=1}^{r} K_{ab}\phi_b\right](E_a + E_{-a}), \tag{1.3}$$

where p_a $[a = 1,2,...,r\ (= \text{rank } G)]$ is the canonical momentum conjugate to ϕ_a; H_a are the Cartan subalgebra generators in the Chevalley basis of the group G, whose Cartan matrix is K_{ab}; E_a (E_{-a}) are the step operators for the simple roots (their negatives) of G; and g_{ab}^{-1} is the inverse of the matrix defined in (2.3a).

The operator $\mathbb{P}$ lives in the tensor product space of two copies of the Lie algebra of G, and is given by[2,7]

$$\mathbb{P} = \sum_{\alpha>0} \frac{\alpha^2}{2}(E_\alpha \otimes E_{-\alpha} - E_{-\alpha} \otimes E_\alpha), \tag{1.4}$$

where the summation is over the positive roots α of G and E_α are the corresponding step operators. The structures of the classical Yang–Baxter equations, like (1.2), are intrinsically related to the algebraic and geometric properties of symmetric spaces.[11]

From (1.2) we obtain

$$\{A, (1/N)\text{Tr } A^N\}_{PB} = [A, B_N], \tag{1.5}$$

where the operator B_N is defined as

$$B_N = \text{Tr}_R[\mathbb{P}(1 \otimes A^{N-1})] = -\text{Tr}_L[\mathbb{P}(A^{N-1} \otimes 1)]. \tag{1.6}$$

The subindices R and L mean we are taking the trace of the right and left entries, respectively, of the tensor product.

In the case where the Hamiltonian is $\text{Tr } A^N/N$, the relation (1.5) becomes a Lax pair equation:[12]

$$\frac{dA}{dt} = [A, B_N]. \tag{1.7}$$

For $N = 2$, this is the Lax equation for the TM equation introduced in (1.1).[2,3] In this sense, the classical Yang–Baxter equation (1.4) is more fundamental than the Lax equation.

From (1.5) it follows that the charges $\text{Tr } A^N$ are in involution, i.e.,

$$\{\text{Tr } A^N, \text{Tr } A^M\}_{PB} = 0. \tag{1.8}$$

 0022-2488/90/123041-06$03.00

The number of these charges, which are functionally independent, is equal to the rank of G.

In the cases where the Hamiltonian is one of those charges, the corresponding system is integrable, since Eq. (1.8) implies it has rank G independent conserved charges (including the Hamiltonian). Therefore, we have rank G integrable systems associated with each classical Yang–Baxter equation (1.2). These constitute the Toda models' hierarchies.[13] The Hamiltonian for the TM model (1.1) is $\mathrm{Tr}\, A^2/2$. It is the simplest model of the hierarchy associated with a given group G.

In this paper, we generalize the method of Refs. 4 and 5 to construct the solutions to higher Hamiltonians in the Toda hierarchy, i.e., the Hamiltonians $\mathrm{Tr}\, A^N/N$ $(N>2)$, which are not quadratic in the momenta. Although these Hamiltonians are representation dependent, our method works uniformly in any representation. The integration of the equations of motion is performed by making a suitable modification of the Lax operator A such that it becomes a polynomial in the momenta of the same degree as B_N. In fact, in a given representation D^λ, we take it to be the component of the matrix $[D^\lambda(A)]^{N-1}$ lying in the representation itself.

In Sec. II we describe the construction of the solutions to the higher Hamiltonians using the zero curvature condition (Lax equation) and the Iwasawa decomposition of the normal real form G^N of G. In Secs. III and IV we apply such a method for the Toda hierarchies associated with the groups SL(3) and SL(4), respectively.

II. THE CONSTRUCTION OF THE SOLUTIONS

Olshanetsky and Perelomov[5] have constructed the solutions of the TM equations (1.1) by projecting, on the phase space of TM, some special geodesic flows on the symmetric phase G^N/K. This space has some nice algebraic properties due to the Iwasawa decomposition[14] of G^N, the (noncompact) normal real form of the Lie group G, whose Cartan matrix appears in (1.1). According to that, the elements of the group G^N decompose as

$$g = nak,\quad g\in G^N, \tag{2.1}$$

where n is an element of the nilpotent subgroup G^N generated by the negative root step operators $E_{-\alpha}$ of G^N, a is an element of the Abelian subgroup generated by the Cartan subalgebra generators H_a $(a=1,2,\ldots,\text{rank } G^N)$, and k belongs to K, which is the maximal compact subgroup of G^N and is generated by $(E_\alpha - E_{-\alpha})$. These generators satisfy the commutation relations

$$[H_a,H_b]=0,\quad a,b=1,2,\ldots,r=\text{rank } G^N, \tag{2.2a}$$

$$[H_a,E_{\pm\alpha}] = \pm K_{a\alpha}E_{\pm\alpha}, \tag{2.2b}$$

$$[E_\alpha,E_\beta] = \begin{cases} N_{\alpha\beta}E_{\alpha+\beta}, & \text{if } \alpha+\beta \text{ is a root},\\ 2\alpha.H/\alpha^2, & \text{if } \alpha+\beta=0,\\ 0, & \text{otherwise}, \end{cases} \tag{2.2c}$$

where $K_{a\alpha}=2\alpha.\alpha_a/\alpha_a^2$, α_a are the simple roots of G^N, and $N_{\alpha\beta}$ are some structure constants that are not important in what follows. The Killing form of G^N can be normalized as

$$\mathrm{Tr}(H_aH_b)=4(\alpha_a.\alpha_b/\alpha_a^2\alpha_b^2)\equiv 4g_{ab}, \tag{2.3a}$$

$$\mathrm{Tr}(E_\alpha H_a)=0, \tag{2.3b}$$

$$\mathrm{Tr}(E_\alpha E_{-\beta})=(2/\alpha^2)\delta_{\alpha\beta}. \tag{2.3c}$$

Due to the Iwasawa decomposition (2.1) of G^N, the points of the symmetric space G^N/K can be put into a one-to-one correspondence with the elements of the solvable subgroup na.[7] This subgroup plays an important role in the construction of the solutions to the Hamiltonians in the Toda hierarchies, as we will now explain.

Although we are dealing with a one-dimensional theory, i.e., which depends only on time, it is much easier to consider a zero curvature condition in two space-time dimensions where the "gauge" potentials, which we shall denote by A_x and A_t, do not depend upon the extra space variable x, i.e., (see comments at the end of this section),[15]

$$[\partial_t + A_t, \partial_x + A_x] = \partial_t A_x - [A_x,A_t] = 0. \tag{2.4}$$

The potentials A_x and A_t will be chosen to be functions of the operators A and B_N, defined in (1.3) and (1.6), respectively, such that the vanishing of (2.4) is a consequence of the Lax pair equation (1.7), and therefore of the equations of motion.

It is highly desirable in our construction, the reason for which will become clear later on, to write, in the light cone variables $u=(x+t)/2$ and $v=(x-t)/2$, the component $A_v = A_x - A_t$ as a "pure gauge" potential of the solvable subgroup na, i.e.,

$$b^{-1}\frac{\partial b}{\partial v} = -b^{-1}\frac{db}{dt} = A_v,\quad b\in na, \tag{2.5}$$

where we have used the fact that the elements b of the subgroup na do not depend upon the x variable. Integrating (2.5), one obtains $b(t)$ as a path ordered integral

$$b(t) = b(0)P\exp\left[\int_0^{-v} A_v\, dv\right]. \tag{2.6}$$

Due to (2.4), the integration above is path independent and therefore we can first integrate in x from 0 to $-t$ and then in t from 0 to t. We get[15]

$$b(t) = b(0)\exp(-tA_x(0))U(t), \tag{2.7}$$

where

$$U(t) = P\exp\left[\int_0^t A_t\, dt\right]. \tag{2.8}$$

So, the unknown time dependence of b is contained in the operator $U(t)$.

We now have to express the potentials A_x and A_t in terms of the operators A and B_N such that A_v is an element of the Lie algebra of na. Using (1.4) and (1.6), we find that B_N is a linear combination of the positive and negative step operators

$$B_N = \sum_{\alpha>0}\frac{\alpha^2}{2}\{\mathrm{Tr}(E_{-\alpha}A^{N-1})E_\alpha - \mathrm{Tr}(E_\alpha A^{N-1})E_{-\alpha}\}. \tag{2.9}$$

For $N=2$, we find using (2.3) and (1.3), that the coefficients of E_α and $E_{-\alpha}$ vanish for α nonsimple, and are the same as the coefficients of $(E_\alpha + E_{-\alpha})$ in (1.3). Therefore, taking $A_x = A$ and $A_t = B_2$, we find that A_v is an element of

the algebra of the solvable subgroup na. This is in fact the "pure gauge" potential used in the construction of the solutions of the TM equations.[4,5,15]

For $N>2$, the situation is more delicate, in part because it involves traces of more than two generators. Since B_N is a polynomial in the momenta of degree $N-1$, we need a potential of the same form to cancel the term proportional to E_α in (2.9) and get a potential in the Lie algebra of na.

In any highest weight finite-dimensional representation D^λ of the Lie algebra g^N of G^N, the generators H_a and E_α satisfy the Hermiticity property

$$D^\lambda(H_a)^+ = D^\lambda(H_a), \quad D^\lambda(E_\alpha)^+ = D^\lambda(E_{-\alpha}). \tag{2.10}$$

It then follows that the Lax operator A, defined in (1.3), is Hermitian:

$$D^\lambda(A)^+ = D^\lambda(A). \tag{2.11}$$

Therefore, in any representation D^λ (of dimension m), the matrix $[D^\lambda(A)]^{N-1}$ is Hermitian and consequently belongs to the vector space V spanned by all $m\times m$ Hermitian matrices [i.e., the defining representation of the Lie algebra of $U(m)$]. We denote by D the subspace of V spanned by $D^\lambda(H_a)$, $D^\lambda(E_\alpha+E_{-\alpha})$, and $iD^\lambda(E_\alpha-E_{-\alpha})$. Since V is a Euclidean vector space, we define M to be the orthogonal complement of D in V. Then

$$\mathrm{Tr}(MD)=0. \tag{2.12}$$

Using the fact that D is a subalgebra of V under the commutator, and the cyclic property of the trace,

$$\mathrm{Tr}\Big(([M,D]D\Big)=\mathrm{Tr}\Big(M[D,D]\Big)=0.$$

Therefore,

$$[M,D]\subset M. \tag{2.13}$$

The matrix $[D^\lambda(A)]^{N-1}$ can be written as a real linear combination of the basis of D plus some matrix M_N^λ, which belongs to M. Then, using (2.3) and (2.12),

$$\begin{aligned} d^\lambda\{[D^\lambda(A)]^{N-1}-M_N^\lambda\} &= \frac{1}{4}\sum_{a,b} g_{ab}^{-1}\,\mathrm{Tr}[D^\lambda(H_bA^{N-1})]D^\lambda(H_a) \\ &\quad+\sum_{\alpha>0}\frac{\alpha^2}{2}\{\mathrm{Tr}[D^\lambda(E_{-\alpha}A^{N-1})]D^\lambda(E_\alpha) \\ &\quad+\mathrm{Tr}[D^\lambda(E_\alpha A^{N-1})]D^\lambda(E_{-\alpha})\}, \end{aligned} \tag{2.14}$$

where d^λ is the Dynkin index[16] of the representation D^λ [the bilinear trace form in D^λ is d^λ times the Killing form (2.3)].

If the Hamiltonian $\mathscr{H}$ is taken to be a function of the charges

$$I_N^\lambda = (1/N)\mathrm{Tr}[D^\lambda(A)]^N, \tag{2.15}$$

then the time evolution of any function f of the canonical variables is given by

$$\frac{df}{dt}=\{f,\mathscr{H}\}_{\mathrm{PB}}=\sum_N\frac{\partial\mathscr{H}}{\partial I_N^\lambda}\{f,I_N^\lambda\}_{\mathrm{PB}}, \tag{2.16}$$

where the summation is over the charges I_N^λ, which are functionally independent. As a consequence of (1.8), the charges I_N^λ are conserved and the Hamiltonian $\mathscr{H}$ is integrable.

From (1.5) and (2.13), we find that the potentials

$$A_x^\lambda(\mathscr{H}) = d^\lambda\sum_N\frac{\partial\mathscr{H}}{\partial I_N^\lambda}\{[D^\lambda(A)]^{N-1}-M_N^\lambda\}, \tag{2.17a}$$

$$A_t^\lambda(\mathscr{H}) = \sum_N\frac{\partial\mathscr{H}}{\partial I_N^\lambda}D^\lambda(B_N) \tag{2.17b}$$

satisfy the Lax pair equation

$$\frac{dA_x^\lambda(\mathscr{H})}{dt}=[A_x^\lambda(\mathscr{H}),A_t^\lambda(\mathscr{H})] \tag{2.18}$$

and, in addition, the matrix M_N^λ has to satisfy

$$\frac{dM_N^\lambda}{dt}-[M_N^\lambda,A_t^\lambda]=0. \tag{2.19}$$

Therefore, traces of powers of M_N^λ are constants of motion.

The advantages of the Lax equation (2.18) are that the potentials A_x^λ and A_t^λ are both polynomials in the momenta of the same degree and, most important, that the potential $A_v^\lambda=A_x^\lambda-A_t^\lambda$ is an element of the Lie algebra of the subgroup na. Therefore, we can take the potentials appearing in (2.4) to be those given by (2.17). From (2.16) and (1.3), we have

$$\frac{d\phi_a}{dt}=\frac{1}{2}\sum_N\frac{\partial\mathscr{H}}{\partial I_N^\lambda}\sum_b g_{ab}^{-1}\,\mathrm{Tr}[D^\lambda(H_bA^{N-1})]. \tag{2.20}$$

Then, using, (2.5), (2.17), (2.14), and (2.9), we have (with b evaluated in the representation D^λ)

$$\begin{aligned} A_v^\lambda(\mathscr{H}) &= -b^{-1}\frac{db}{dt} \\ &=\frac{1}{2}\sum_a\frac{d\phi_a}{dt}D^\lambda(H_a)+2\sum_N\frac{\partial\mathscr{H}}{\partial I_N^\lambda} \\ &\quad\times\sum_{\alpha>0}\frac{\alpha^2}{2}\mathrm{Tr}[D^\lambda(E_\alpha A^{N-1})]D^\lambda(E_{-\alpha}). \end{aligned} \tag{2.21}$$

Therefore, the coefficients of the Cartan subalgebra generators in A_v^λ are just the velocities. This fact makes the integration of the equations of motion easier. Indeed, since $b=na$, we have

$$b^{-1}\frac{db}{dt}=a^{-1}\frac{da}{dt}+a^{-1}n^{-1}\frac{dn}{dt}a. \tag{2.22}$$

The first term on the rhs is a linear combination of the Cartan subalgebra generators, and the second term is a linear combination of negative step operators. Therefore, comparing (2.21) with (2.22), we get

$$a^{-1}\frac{da}{dt}=-\frac{1}{2}\sum_a\frac{d\phi_a}{dt}D^\lambda(H_a) \tag{2.23}$$

and, consequently,

$$a=\exp(-\tfrac{1}{2}\phi^\lambda(t)), \tag{2.24}$$

where we have defined

$$\phi^\lambda(t)=\sum_a\phi_a(t)D^\lambda(H_a). \tag{2.25}$$

Using (2.10) and (2.11), we have

$$\mathrm{Tr}\, D^{\lambda}(H_a A^{N-1})^{*} = \mathrm{Tr}\, D^{\lambda}(H_a A^{N-1}),$$
$$\mathrm{Tr}\, D^{\lambda}(E_a A^{N-1})^{*} = \mathrm{Tr}\, D^{\lambda}(E_{-a} A^{N-1}). \tag{2.26}$$

From (2.9), we then see that the operator B_N is anti-Hermitian, and therefore the potentials introduced in (2.17) satisfy

$$[A_x^{\lambda}(\mathscr{H})]^{+} = A_x^{\lambda}(\mathscr{H}), \quad [A_t^{\lambda}(\mathscr{H})]^{+} = -A_t^{\lambda}(\mathscr{H}). \tag{2.27}$$

Working with these potentials, we find that the operator $U(t)$, defined in (2.8), is unitary and consequently we have, from (2.7),

$$b(t)b(t)^{+} = b(0)\exp\{-2tA_x^{\lambda}(\mathscr{H})|_{t=0}\}b(0)^{+}. \tag{2.28}$$

Using (2.24) and the fact that $\phi^{\lambda}(t)$ is Hermitian, we get

$$e^{-\phi^{\lambda}(t)} = n(t)^{-1}n(0)e^{-\phi^{\lambda}(0)/2} \times\exp\{-2tA_x^{\lambda}(\mathscr{H})|_{t=0}\} \times e^{-\phi^{\lambda}(0)/2}n(0)^{+}[n(t)^{+}]^{-1}. \tag{2.29}$$

Therefore, the unknown time dependence of the coordinates ϕ_a is contained in the elements of the nilpotent subgroup n, whose generators are the negative step operators. However, by taking the expectation value of the expression (2.29) in the highest weight state $|\lambda\rangle$ of the representation D^{λ}, we eliminate this unknown time dependence. Indeed, the state $|\lambda\rangle$ is annihilated by all positive step operators, and consequently

$$n^{+}|\lambda\rangle = |\lambda\rangle. \tag{2.30}$$

The fundamental weights λ_a ($a = 1,2,\ldots,\text{rank } G^N$) satisfy

$$2\alpha_a.\lambda_b/\alpha_a^2 = \delta_{ab}. \tag{2.31}$$

Therefore, working with the rank G^N fundamental representations of G^N, we get

$$D^{\lambda_a}(H_b)|\lambda_a\rangle = \delta_{ab}|\lambda_a\rangle. \tag{2.32}$$

Consequently, from (2.29) and (2.30), it follows that

$$e^{-\phi_a(t)} = e^{-\phi_a(0)}\langle\lambda_a|\exp[-2tA_x^{\lambda_a}(\mathscr{H})|_{t=0}]|\lambda_a\rangle. \tag{2.33}$$

This is the general solution of the equations of motion for a model described by any Hamiltonian $\mathscr{H}$, which is a function of the charges $\mathrm{Tr}\, A^N$. We notice that the solutions have the same formal expression for any $\mathscr{H}$, including the solutions to the Toda molecule models[4,5] corresponding to $\mathscr{H} = (1/2)\mathrm{Tr}\, A^2$. The actual different time evolutions are encoded into the operator $A_x^{\lambda_a}(\mathscr{H})|_{t=0}$ containing the initial values of coordinates and momenta.

The introduction of the extra space varaiable x is an artifice of the construction. It enables us to treat the Lax operators as the components of a "gauge potential" in two dimensions. Then we can play with the path-independent integration of (2.5) to select a component of the gauge potential, which points to a suitable direction in the algebra g^N. In fact, the choice of A_x^{λ} is made in such a way that A_v^{λ} lies in the subalgebra na. This choice implies, in addition, that the coefficients of the Cartan subalgebra generators in A_v^{λ} are just the velocities. This makes the integration possible. We point out, however, that the physics of the problem is unchanged by the introduction of the extra space variable x since the gauge potentials are x independent.

In the next two sections, we show how to use the procedure described above to find the general solutions to the Hamiltonians in the hierarchies of the Toda models associated with the groups SL(3) and SL(4).

III. THE SL(3) TODA HIERARCHY

For the case of SL(3), there are only two charges $\mathrm{Tr}\, A^N$ that are functionally independent, namely, $\mathrm{Tr}\, A^2$ and $\mathrm{Tr}\, A^3$. In the case where the Hamiltonian is $\mathrm{Tr}\, A^2$, we get the Toda molecule equations for SL(3), whose solutions are known.[4,5] We now want to discuss the model defined by the Hamiltonian $\mathrm{Tr}\, A^3$.

The equations become simpler if we perform the canonical transformation

$$\rho_a = 2\sum_b g_{ab}\phi_b = \sum_b K_{ab}\frac{\phi_b}{\alpha_a^2}, \tag{3.1a}$$

$$\pi_a = \frac{1}{2}\sum_b g_{ab}^{-1}p_b. \tag{3.1b}$$

The Lax operator A, defined in (1.3), then becomes

$$A = \sum_a \pi_a H_a + \frac{1}{2}\sum_a \exp\left(\rho_a\frac{\alpha_a^2}{2}\right)(E_a + E_{-a}). \tag{3.2}$$

All roots of SL(3) have the same length and so we can set $\alpha_a^2 = 2$. We shall denote by λ_1 and λ_2 the fundamental weights of SL(3), which are associated to the triplet and antitriplet representations, respectively. By taking the Hamiltonian $\mathscr{H}$ of our system to be $\mathrm{Tr}\, A^3/3$, evaluated in the triplet representation, we get

$$\mathscr{H} = \tfrac{1}{3}\mathrm{Tr}[D^{\lambda_1}(A)]^3 = \pi_1\pi_2(\pi_1 - \pi_2) + \tfrac{1}{4}[\pi_2 e^{2\rho_1} - \pi_1 e^{2\rho_2}]. \tag{3.3}$$

When evaluated in the antitriplet representation, the quantity $\mathrm{Tr}\, A^3$ gets a factor (-1) with respect to its value in the triplet representation. So we have

$$\mathscr{H} = -\tfrac{1}{3}\mathrm{Tr}[D^{\lambda_2}(A)]^3. \tag{3.4}$$

This Hamiltonian is not positive definite and it is singular whenever

$$\pi_1 = e^{\pm i\pi/3}\pi_2. \tag{3.5}$$

However, it describes an integrable system since, from (1.8), we get that the quantity

$$I_2 = \tfrac{1}{2}\mathrm{Tr}[D^{\lambda_1}(A)]^2 = \tfrac{1}{2}\mathrm{Tr}[D^{\lambda_2}(A)]^2 = \pi_1^2 + \pi_2^2 - \pi_1\pi_2 + \tfrac{1}{4}[e^{2\rho_1} + e^{2\rho_2}] \tag{3.6}$$

is conserved.

The Hamilton's equations of motion for such systems are given by

$$\dot{\rho}_1 = \pi_2(2\pi_1 - \pi_2) - \tfrac{1}{4}e^{2\rho_2}, \tag{3.7a}$$

$$\dot{\rho}_2 = -\pi_1(2\pi_2 - \pi_1) + \tfrac{1}{4}e^{2\rho_1}, \tag{3.7b}$$

$$\dot{\pi}_1 = -\tfrac{1}{2}\pi_2 e^{2\rho_1}, \tag{3.7c}$$

$$\dot{\pi}_2 = \tfrac{1}{2}\pi_1 e^{2\rho_2}. \tag{3.7d}$$

Notice that if the momenta vanish at a given time, then they will vanish for all times. In this case the energy of the system is zero. According to (3.5) this corresponds to a singular point of the Hamiltonian. The dynamics of the system at this singular point become very simple since Eqs. (3.7) reduce to

$$\ddot{\rho}_- = 0, \tag{3.8a}$$

$$\ddot{\rho}_+ = -\tfrac{1}{4} e^{2\rho_+}, \tag{3.8b}$$

where $\rho_\pm = \rho_1 \pm \rho_2$. Therefore, in this case the system is composed of a particle satisfying the Liouville equation plus a free particle. However, due to the conservation of I_2, they are subjected to the constraint

$$\tfrac{1}{2}\dot{\rho}_-^2 = \tfrac{1}{2}\dot{\rho}_+^2 + \tfrac{1}{8} e^{2\rho_+}. \tag{3.9}$$

The equations of motion (3.7) can be written in a more interesting form by eliminating the momenta using the conserved quantity I_2. From (3.7) we have

$$2\ddot{\rho}_1 + \ddot{\rho}_2 = -(\pi_2^2 + \pi_1\pi_2 + \tfrac{1}{4}e^{2\rho_2} - \tfrac{1}{2}\dot{\rho}_1)e^{2\rho_1}, \tag{3.10a}$$

$$2\ddot{\rho}_2 + \ddot{\rho}_1 = -(\pi_1^2 + \pi_1\pi_2 + \tfrac{1}{4}e^{2\rho_1} + \tfrac{1}{2}\dot{\rho}_2)e^{2\rho_2}. \tag{3.10b}$$

Using (3.1), (3.6), (3.7a), (3.7b), and the fact that, for SL(3), $K_{11} = K_{22} = 2$, $K_{12} = K_{21} = -1$, we can write (3.10) as

$$\ddot{\phi}_a = \left[\tfrac{1}{2}\epsilon_{ac}\dot{\phi}_c - \tfrac{2}{3}I_2\right]e^{K_{ab}\phi_b}, \quad a = 1,2, \tag{3.11}$$

where $\epsilon_{11} = \epsilon_{22} = 0$, $\epsilon_{12} = -\epsilon_{21} = 1$.

We now construct the solutions to this equation using the procedure of Sec. II. In the l-dimensional fundamental representations of SL(l), the matrix M_N^λ is proportional to the $l \times l$ unity matrix. The reason is that any $l \times l$ real matrix can be written as a linear combination of the matrices representing the $(l^2 - 1)$ generators of SL(l), in one of these representations, and the unity matrix. The trace of a generator of a semisimple Lie algebra vanishes in any finite-dimensional representation. Then, by taking the trace on both sides of (2.14) and using (3.6), we get

$$M_3^{\lambda_1} = M_3^{\lambda_2} = \tfrac{2}{3}I_2 \mathbb{1}_{3\times 3}, \tag{3.12}$$

which obviously satisfies (2.19).

Therefore, the potentials (2.17), in the triplet representation, are given by

$$A_x^{\lambda_1} = [D^{\lambda_1}(A)]^2 - \tfrac{2}{3}I_2\mathbb{1}_{3\times 3}, \tag{3.13a}$$

$$A_t^{\lambda_1} = D^{\lambda_1}(B_3), \tag{3.13b}$$

and in the antitriplet representation by

$$A_x^{\lambda_2} = -[D^{\lambda_2}(A)]^2 + \tfrac{2}{3}I_2\mathbb{1}_{3\times 3}, \tag{3.14a}$$

$$A_t^{\lambda_2} = -D^{\lambda_2}(B_3), \tag{3.14b}$$

where we have used the fact that $d^{\lambda_1} = d^{\lambda_2} = 1$.

Since λ_1 and λ_2 are the fundamental weights of SL(3), the expression for the solutions $\phi_1(t)$ and $\phi_2(t)$ are given by (2.33) for the potentials (3.13a) and (3.14a), respectively.

IV. THE SL(4) TODA HIERARCHY

We will denote by λ_1, λ_2, and λ_3 the fundamental weights of SL(4). They are, respectively, the highest weights of the 4, 6, and $\bar{4}$ fundamental representations of SL(4). In this case there are three charges Tr A^N that are functionally independent. Using the notation introduced in (2.15), we find that in the D^{λ_1} fundamental representation these three charges are

$$I_2^{\lambda_1} = \pi_1^2 + \pi_2^2 + \pi_3^2 - \pi_1\pi_2 - \pi_2\pi_3 + \tfrac{1}{4}(e^{2\rho_1} + e^{2\rho_2} + e^{2\rho_3}), \tag{4.1a}$$

$$I_3^{\lambda_1} = \pi_1\pi_2(\pi_1 - \pi_2) + \pi_2\pi_3(\pi_2 - \pi_3) + \tfrac{1}{4}[\pi_2 e^{2\rho_1} + (\pi_3 - \pi_1)e^{2\rho_2} - \pi_2 e^{2\rho_3}], \tag{4.1b}$$

$$\begin{aligned} I_4^{\lambda_1} = {} & \tfrac{1}{2}\{\pi_1^4 + \pi_2^4 + \pi_3^4 - 2\pi_2(\pi_1^3 + \pi_3^3) + 3\pi_2^2(\pi_1^2 + \pi_3^2) \\ & - 2\pi_2^3(\pi_1 + \pi_3) + \tfrac{1}{2}(\pi_1^2 - \pi_1\pi_2 + \pi_2^2)e^{2\rho_1} \\ & + \tfrac{1}{2}(\pi_3^2 - \pi_2\pi_3 + \pi_2^2)e^{2\rho_3} \\ & + \tfrac{1}{2}[\pi_1^2 + \pi_2^2 + \pi_3^2 - \pi_2(\pi_1 + \pi_3) - \pi_1\pi_3]e^{2\rho_2} \\ & + \tfrac{1}{16}(e^{4\rho_1} + e^{4\rho_2} + e^{4\rho_3} + 2e^{2(\rho_1+\rho_2)} + 2e^{2(\rho_2+\rho_3)})\}, \end{aligned} \tag{4.1c}$$

where we have used the canonical variables introduced in (3.1).

In the D^{λ_3} fundamental representation we have

$$I_2^{\lambda_3} = I_2^{\lambda_1}, \quad I_3^{\lambda_3} = -I_3^{\lambda_1}, \quad I_4^{\lambda_3} = I_4^{\lambda_4}. \tag{4.2}$$

In the D^{λ_2} fundamental representation the charges $I_3^{\lambda_2}$ and $I_5^{\lambda_2}$ vanish. Then the functionally independent charges in this representation are

$$I_2^{\lambda_2} = 2I_2^{\lambda_1}, \tag{4.3a}$$

$$I_4^{\lambda_2} = 3(I_2^{\lambda_1})^2 - 4I_4^{\lambda_1}, \tag{4.3b}$$

$$I_6^{\lambda_2} = \tfrac{14}{3}(I_2^{\lambda_1})^3 - 8I_2^{\lambda_1}I_4^{\lambda_1} + (I_3^{\lambda_1})^2. \tag{4.3c}$$

The relations (4.1)–(4.3) can be easily checked using a REDUCE program.

The solutions to any Hamiltonian that is a function of the charges Tr A^N can be written in terms of the solutions of three Hamiltonians that are independent functions of these charges. The SL(4) Toda molecule model itself is described by the Hamiltonian $I_2^{\lambda_1}$. Therefore, we will take the Hamiltonians of the other two models in this hierarchy to be

$$\mathscr{H}_4 = I_4^{\lambda_2}, \tag{4.4}$$

$$\mathscr{H}_6 = I_6^{\lambda_2}. \tag{4.5}$$

We have

$$M_2^{\lambda_1} = 0, \quad M_3^{\lambda_1} = \tfrac{1}{2}I_2^{\lambda_1}\mathbb{1}_{4\times 4}, \quad M_4^{\lambda_1} = \tfrac{3}{4}I_3^{\lambda_1}\mathbb{1}_{4\times 4}, \tag{4.6}$$

$$M_2^{\lambda_2} = M_4^{\lambda_2} = M_6^{\lambda_2} = 0, \tag{4.7}$$

$$M_2^{\lambda_3} = 0, \quad M_3^{\lambda_3} = \tfrac{1}{2}I_2^{\lambda_3}\mathbb{1}_{4\times 4}, \quad M_4^{\lambda_3} = \tfrac{3}{4}I_3^{\lambda_3}\mathbb{1}_{4\times 4}. \tag{4.8}$$

Equations (4.6) and (4.8) are obtained using the same arguments leading to (3.12). Equation (4.7) can be checked using a REDUCE program. The Dynkin indices for the fundamental representations of SL(4) are $d^{\lambda_1} = d^{\lambda_3} = 1$ and $d^{\lambda_2} = 2$. Therefore, from (2.17), (4.2), and (4.3), we find that the Lax pair operators for the Hamiltonian $\mathscr{H}_4$, in the fundamental representations, are

$$A_x^{\lambda_1}(\mathscr{H}_4) = 6I_2^{\lambda_1}D^{\lambda_1}(A) - 4\{[D^{\lambda_1}(A)]^3 - \tfrac{3}{4}I_3^{\lambda_1}\mathbb{1}_{4\times 4}\}, \tag{4.9a}$$

$$A_t^{\lambda_1}(\mathscr{H}_4) = 6I_2^{\lambda_1}D^{\lambda_1}(B_2) - 4D^{\lambda_1}(B_4), \tag{4.9b}$$

$$A_x^{\lambda_2}(\mathscr{H}_4) = 2[D^{\lambda_2}(A)]^3, \tag{4.10a}$$

$$A_t^{\lambda_2}(\mathscr{H}_4) = D^{\lambda_2}(B_4), \tag{4.10b}$$

$$A_x^{\lambda_3}(\mathscr{H}_4) = 6I_2^{\lambda_3}D^{\lambda_3}(A) - 4\{[D^{\lambda_3}(A)]^3 - \tfrac{3}{4}I_3^{\lambda_3}\mathbb{1}_{4\times 4}\}, \tag{4.11a}$$

$$A_t^{\lambda_3}(\mathscr{H}_4) = 6I_2^{\lambda_3}D^{\lambda_3}(B_2) - 4D^{\lambda_3}(B_4). \tag{4.11b}$$

Analogously, the Lax pair operators for the Hamiltonian $\mathscr{H}_6$ in the fundamental representations are

$$\begin{aligned} A_x^{\lambda_1}(\mathscr{H}_6) = {} & [14(I_2^{\lambda_1})^2 - 8I_4^{\lambda_1}]D^{\lambda_1}(A) \\ & + 2I_3^{\lambda_1}\{[D^{\lambda_1}(A)]^2 - \tfrac{1}{2}I_2^{\lambda_1}\mathbb{1}_{4\times 4}\} \\ & - 8I_2^{\lambda_1}\{[D^{\lambda_1}(A)]^3 - \tfrac{3}{4}I_3^{\lambda_1}\mathbb{1}_{4\times 4}\}, \end{aligned} \tag{4.12a}$$

$$\begin{aligned} A_t^{\lambda_1}(\mathscr{H}_6) = {} & [14(I_2^{\lambda_1})^2 - 8I_4^{\lambda_1}]D^{\lambda_1}(B_2) \\ & + 2I_3^{\lambda_1}D^{\lambda_1}(B_3) - 8I_2^{\lambda_1}D^{\lambda_1}(B_4), \end{aligned} \tag{4.12b}$$

$$A_x^{\lambda_2}(\mathscr{H}_6) = 2[D^{\lambda_2}(A)]^5, \tag{4.13a}$$

$$A_t^{\lambda_2}(\mathscr{H}_6) = D^{\lambda_2}(B_6), \tag{4.13b}$$

$$\begin{aligned} A_x^{\lambda_3}(\mathscr{H}_6) = {} & [14(I_2^{\lambda_3})^2 - 8I_4^{\lambda_1}]D^{\lambda_3}(A) \\ & + 2I_3^{\lambda_3}\{[D^{\lambda_3}(A)]^2 - \tfrac{1}{2}I_2^{\lambda_3}\mathbb{1}_{4\times 4}\} \\ & - 8I_2^{\lambda_3}\{[D^{\lambda_3}(A)]^3 - \tfrac{3}{4}I_3^{\lambda_3}\mathbb{1}_{4\times 4}\}, \end{aligned} \tag{4.14a}$$

$$\begin{aligned} A_t^{\lambda_3}(\mathscr{H}_6) = {} & [14(I_2^{\lambda_3})^2 - 8I_4^{\lambda_1}]D^{\lambda_3}(B_2) \\ & + 2I_3^{\lambda_3}D^{\lambda_3}(B_3) - 8I_2^{\lambda_3}D^{\lambda_3}(B_4). \end{aligned} \tag{4.14b}$$

The general solutions to the Hamiltonians $\mathscr{H}_4$ and $\mathscr{H}_6$ are obtained from (2.33) using the "gauge potentials" given above.

V. CONCLUSIONS

We have presented a method for constructing the general solution to higher Hamiltonians in the Toda hierarchies associated to any simple Lie group G. The method is representation independent and it is a generalization of the work of Olshanetsky and Perelomov,[5] and Leznov and Saveliev.[4] The key point of the construction is the modification of the Lax operator. For the Toda molecule models the operators A and B_2, defined in (1.3) and (1.6), respectively, are such that their difference is an element of the solvable subalgebra na. This fact plays an important role in the construction of the solutions of the TM[4,5] since it makes the connection with the geodesics on the symmetric space G^N/K.[5,7] For the higher Hamiltonians, $\mathrm{Tr}\,A^N$ $(N>2)$, in the hierarchies the operator B_N depends upon the momenta and, unlike A, it contains nonsimple root step operators. The operator A is Hermitian in any representation D^λ, and therefore powers of the matrix $D^\lambda(A)$ (of dimension m) belong to the vector space of the $m\times m$ Hermitian matrices. Since this is a Euclidean space, the matrices belonging to the representation D^λ can be split from the rest in a way that relations (2.12) and (2.13) hold. Then, by replacing the operator A by the component of the matrix $[D^\lambda(A)]^{N-1}$ lying in the representation D^λ, we obtain from (1.7) two decoupled Lax equations, namely (2.18) and (2.19). The first one leads us to the solution in a quite simple way. Analogously to the TM case, the difference between the new Lax operators (2.17) is an element of the subalgebra na. In addition, the integration is made easy by the fact the component of the Lax operator (2.21) lying in the Cartan subalgebra is linear in the velocities. This is very similar to what happens in Refs. 4 and 5.

It would be very interesting to investigate further the possibility of understanding this construction in terms of the universal enveloping algebra of G^N. We believe that such an investigation could shed some light on the quantum integrability properties of these models and their relation with quantum groups. Natural extensions of our work would be the construction of the solutions of the two-dimensional version of these higher Hamiltonians. One could also try to make a connection between the solution presented here and special motions on the symmetric spaces G^N/K.

ACKNOWLEDGMENTS

We are grateful to J. J. Giambiagi for his support throughout this work, to M. A. Manna for many helpful discussions, and to J. C. Monteiro for helping us with the REDUCE program. One of us (RML) is grateful to IFT-UNESP for their hospitality.

This work was partially supported by CNPq (Brazil).

[1] M. Toda, Phys. Rep. C**18**, 1 (1975).
[2] D. I. Olive and N. Turok, Nucl. Phys. B **220** [FS8], 491 (1983).
[3] O. I. Bogoyavlensky, Commun. Math. Phys. **51**, 201 (1976).
[4] A. M. Leznov and M. V. Saveliev, Lett. Math. Phys. **3**, 489 (1979); Phys. Lett. B**79**, 294 (1978); Commun. Math. Phys. **74**, 111 (1980).
[5] M. A. Olshanetsky and A. M. Perelomov, Inventiones Math. **54**, 261 (1979); Theor. Math. Phys. **45**, 843 (1980).
[6] B. Kostant, Adv. Math. **34**, 195 (1979).
[7] L. A. Ferreira and D. I. Olive, Commun. Math. Phys. **99**, 365 (1985).
[8] L. Faddeev, "Integrable Models in 1 + 1 Dimensional Quantum Field Theory," in *Proceedings of the École D'Été de Physique Théorique*, Les Houches, 1982 (North Holland, Amsterdam, 1982).
[9] R. J. Baxter, Ann. Phys. **70**, 193 (1972); *Exactly Solved Models in Statistical Mechanics* (Academic, New York, 1982).
[10] C. N. Yang, Phys. Rev. Lett. **19**, 1312 (1967).
[11] L. A. Ferreira, Int. J. Mod. Phys. A **4**, 649 (1989); **4**, 675 (1989).
[12] P. Lax, Commun. Pure Appl. Math. **21**, 467 (1968).
[13] D. I. Olive and N. Turok, Nucl. Phys. B **265** [FS15], 469 (1986).
[14] S. Helgason, *Differential Geometry, Lie Groups and Symmetric Spaces* (Academic, New York, 1978).
[15] D. I. Olive, "Lectures on Gauge Theories and Lie Algebras," Univ. of Virginia preprint (1982).
[16] E. B. Dynkin, Trans. Am. Math. Soc. **6** (2), 111 (1957).

Unified prescription for the generation of electroweak and gravitational gauge field Lagrangian on a principal fiber bundle

Theo Verwimp
Physics Departement, U.I.A., Universiteit Antwerpen B-2610 Wilrijk, Antwerpen, Belgium

(Received 7 December 1989; accepted for publication 1 August 1990)

It is essentially known that useful gauge field Lagrangians arise as Weil polynomials of the curvature of the gauge connection. The deeper implications and details of this fact are worked out in two widely differing cases. The Glashow–Weinberg–Salam gauge field Lagrangian for electroweak theory and the Townsend–Zardecki action for gravitation are obtained from the same type of "Yang–Mills" Weil form on a principal fiber bundle over space-time, with symmetry group U(2) and SO(2,3), respectively. The unified geometrical approach given here shows that fiber bundle reduction and symmetry breaking are essential not only in electroweak theory but also in the SO(2,3) gauge theory for gravitation. In fact, the process of symmetry breaking in electroweak theory and the soldering of the anti-de Sitter bundle, essential in the interpretation of SO(2,3) gauge theory as a theory for gravitation, are corresponding geometrical concepts.

I. INTRODUCTION

It has been long known that from the unifying point of view of fiber bundles, the kinematics of gauge theories and general relativity are the same.[1] However, the gauge scheme itself fails to determine directly the dynamics of the theory, i.e., a form of Lagrangian. Euler–Lagrange equations are obtained after a Lagrangian has been defined. For example, in a Yang–Mills theory with a Lie group G, the gauge field Lagrangian is defined as

$$I_{AB}F^{A}{}_{\mu\nu}F^{B\mu\nu}, \tag{1.1}$$

where $F^{A\mu\nu}$ are the components of the curvature tensor $F^{\mu\nu}$ of the gauge field relative to a basis of $\mathcal{G}$, the Lie algebra of G, and I_{AB} is the metric tensor of the bilinear Killing form of G, $A, B = 1,...,k$, k the dimension of $\mathcal{G}$.[2]

Within the fiber bundle formulation of gauge theories it is, however, more natural to define a Lagrangian form on the principal fiber bundle $P(M,G)$ over space-time M. A most general gauge invariant Lagrangian form on P that projects uniquely to the base manifold can be defined using Weil forms[3] constructed from Weil polynomials[4] and $\mathcal{G}$-valued tensorial two-forms on P. Recently, Weil polynomials were used by Kakazu and Matsumoto in the reconstruction of Einstein gravity (with torsion) and four-dimensional $N = 1$ supergravity on a principal fiber bundle (PFB).[5] It was also possible to reconstruct the Lovelock Lagrangian[6] for higher dimensional gravity on the bundle of orthonormal frames over a space-time of arbitrary dimension D by using Weil polynomials on the Lie algebra so$(1,D-1)$.[3]

In this paper, we will emphasize the complete analogy between the derivation of the Glashow–Weinberg–Salam gauge field Lagrangian for electroweak theory and the Townsend–Zardecki type action for gravitation[7] from the same type of Weil form on a PFB with symmetry group U(2) and SO(2,3), respectively. Besides the Einstein action and a cosmological term, the Townsend–Zardecki action also contains curvature-squared and torsion-squared terms. However, for zero torsion and zero effective cosmological constant, the theory reduces essentially to Einstein's theory in vacuum.[7] Both Townsend and Zardecki derive their Lagrangian in the context of a SO(1,4) gauge theory. Here, we will consider as gauge group for gravitation the anti-de Sitter group SO(2,3). This choice is inspired by the proposal of Ward to use the real Lie algebra so*(14), as the basis for a unified gauge theory of elementary particles.[8] In fact,

$$\mathrm{so}(2,3)\oplus\mathrm{su}(3)\oplus\mathrm{su}(2)\oplus\mathrm{u}(1)\subset\mathrm{so}^*(6)$$
$$\oplus\,\mathrm{so}^*(8)\subset\mathrm{so}^*(14)\subset\mathrm{so}^*(16) \tag{1.2}$$

and so*(16) is a maximum subalgebra of a noncompact real form[9] of the exceptional Lie algebra E_8. Much of the formalism exposed in this paper can be useful in the construction of a gauge field Lagrangian for a unified theory of electroweak, strong, and gravitational interaction based on (1.2).

The rest of the paper is organized as follows. In Sec. II we describe the reduction of a principal fiber bundle P to a subbundle Q, the splitting of a connection one-form on P into a reduced connection on Q and a tensorial one-form, and give the expressions for the reduced curvature on Q. The occurrence of reduction is closely related with the process of symmetry breaking. In the electroweak theory the remaining symmetry is that of the charge conservation group $\mathrm{U}_c(1)$, while for gravitation we consider symmetry breaking from SO(2,3) to SO(1,3). In Sec. III we outline the construction of a gauge invariant Lagrangian form on the U(2) and the SO(2,3) principal fiber bundle, respectively, and determine the Weil polynomials needed in the construction of these Lagrangian forms. With the theory thus developed it is then straightforward to determine the electroweak and gravitational gauge field Lagrangian in Secs. IV and V. Although the derivation of the gravitational gauge field Lagrangian is, except for the gauge group, not essentially different from that given by Zardecki,[7] our unified geometrical approach, however, will make it possible to identify corresponding fields and concepts in the electroweak and gravitational gauge theories.

 0022-2488/90/123047-06$03.00

II. FIBER BUNDLE REDUCTION

Let $P(M,G)$ be a PFB with base space M and structure group G. If $\rho: G \to \mathrm{GL}(V)$ is a representation of G in a finite-dimensional vector space V, then a pseudotensorial form of degree r on P of type (ρ,V) is an r-form $\psi: P \to V$, which is equivariant with respect to the action of G, i.e.,

$$R_g^* \psi = \rho(g^{-1}) \cdot \psi, \tag{2.1}$$

where R_g is the right action of $g \in G$ on P. Such a form ψ is called a tensorial form if $\psi(X_1,...,X_r) = 0$ whenever at least one of the tangent vectors X_i of P is tangent to a fiber.[4] In particular, a tensorial zero-form of type (ρ,V) is a function $\psi: P \to V$ such that

$$\psi(ug) = \rho(g^{-1}) \cdot \psi(u), \quad \text{for } u \in P,\ g \in G. \tag{2.2}$$

Suppose that H is a closed subgroup of G. On the coset space G/H, the Lie group G acts as a transitive Lie transformation group, i.e., G/H is a homogeneous manifold of G. If $G/H \subset V$ the vector space on which G acts through ρ, then the coset space G/H can be thought of as the orbit space $V_0 = \rho(G) \cdot v_0$ with $v_0 \in G/H$ an H-fixed point in V. Now, let $E(M,G/H,G,P)$ be the vector bundle associated with P and with standard fiber G/H. Then one has the following result:[4] The bundle $P(M,G)$ has a reduction to a subbundle $Q(M,H)$ iff E admits a global section. Moreover, there is a one-to-one correspondence between cross sections $\sigma: M \to E$ and equivariant mappings $\psi: P \to G/H$. Therefore, a necessary and sufficient condition for the occurrence of the reduction is the existence on P of a tensorial zero-form ψ of type (ρ,V) with range $\psi(P)$ in the orbit space G/H (see also Ref. 10). Thereby $Q = \psi^{-1}(v_0)$, and Fulp and Norris in Ref. 11 refer to ψ as a symmetry breaking Higgs field.

If the symmetry G is a gauge symmetry associated with a Lagrangian field theoretical model, then the original symmetry G is said to be broken spontaneously by the Higgs mechanism iff the self-interaction potential $V(\Psi)$ of a multiplet Ψ of (possibly complex) scalar fields, which are described by a global section of a vector bundle W associated with P with fiber $V \supset G/H$, is minimized by the manifold $V_0 = G/H$. In that case, the global section $\Psi_0(x) = v_0$ is called the vacuum or ground state and the homogeneous space V_0, obtained by the action of $\rho(G)$ on v_0, the vacuum manifold.

Let $\gamma: Q \to P$ be the imbedding of the reduced bundle $Q(M,H)$ into its extension $P(M,G)$, and set

$$\mathcal{G} = \gamma_*(\mathcal{H}) \oplus \mathcal{C}, \tag{2.3}$$

where $\mathcal{H}$ is the Lie algebra of the subgroup H and $\mathcal{C}$ a vector subspace of $\mathcal{G}$, the Lie algebra of G. If $\tilde{\mu}$ is a connection form on P, then the restriction μ of $\tilde{\mu}$ to Q, i.e., $\mu = \gamma^* \tilde{\mu}$, splits naturally into

$$\mu = \omega + \phi, \tag{2.4}$$

where ω is $\mathcal{H}$ valued and ϕ is $\mathcal{C}$ valued. We then have the theorem[12] that if the complement $\mathcal{C}$ of $\mathcal{H}$ in $\mathcal{G}$ is Ad invariant by H or equivalently $[\gamma_* \mathcal{H}, \mathcal{C}] \subset \mathcal{C}$ if H is connected, then the restriction ω of the $\mathcal{H}$ component of the connection one-form $\tilde{\mu}$ on P is a connection one-form on Q while ϕ is a tensorial one-form of type $(\mathrm{Ad}\, H, \mathcal{C})$ on Q. Here, Ad always denotes the adjoint representation of the symmetry group on its Lie algebra.

The curvature two-form $\tilde{\Delta}$ on P is a tensorial two-form of type $(\mathrm{Ad}\, G, \mathcal{G})$ defined as

$$\tilde{\Delta} = D\tilde{\mu} = d\tilde{\mu} + \tfrac{1}{2}[\tilde{\mu},\tilde{\mu}], \tag{2.5}$$

with $[,]$ denoting the exterior product of forms with values in a Lie algebra. The reduction $\Delta = \gamma^* \tilde{\Delta}$ to Q of the curvature $\tilde{\Delta}$ calculated from $\tilde{\mu}$ on P can be written[5]

$$\Delta = \Omega + \Phi + \Sigma, \tag{2.6}$$

where

$$\Omega = d\omega + \tfrac{1}{2}[\omega,\omega], \tag{2.7}$$

$$\Phi = d\phi + [\omega,\phi], \tag{2.8}$$

$$\Sigma = \tfrac{1}{2}[\phi,\phi]. \tag{2.9}$$

From the Bianchi identity on P

$$d\tilde{\Delta} + [\tilde{\mu},\tilde{\Delta}] = 0, \tag{2.10}$$

we find after reduction to the subbundle Q,

$$D\Omega = d\Omega + [\omega,\Omega] = 0, \tag{2.11}$$

$$D\Phi = d\Phi + [\omega,\Phi] = [\Omega,\phi]. \tag{2.12}$$

III. A GAUGE INVARIANT LAGRANGIAN FORM

In general, $\mathrm{Ad}(G)$-invariant Weil polynomials[4,5,13] of degree m on the Lie algebra $\mathcal{G}$ are defined as multilinear symmetric real-valued functions L_m such that

$$L_m(g \cdot T_1,...,g \cdot T_m) = L_m(T_1,...,T_m), \tag{3.1}$$

for all $g \in G$, $T_i \in \mathcal{G}$ and where $g \cdot T_i = gT_ig^{-1}$. The space of all Weil polynomials of degree m is denoted by $S_G^m(\mathcal{G})$.

If $\tilde{\psi}_1,...,\tilde{\psi}_m$ are $\mathcal{G}$-valued two-forms on P, then we define for each $L_m \in S_G^m(\mathcal{G})$ a real-valued Weil form $L_m(\tilde{\psi}_1,...,\tilde{\psi}_m)$ on P of degree $2m = D$ the dimension of M, by

$$\begin{aligned} &L_m(\tilde{\psi}_1,...,\tilde{\psi}_m)(X_1,...,X_D) \\ &\quad = \left(\frac{1}{2}\right)^m \sum_{\sigma \in S_D} \epsilon(\sigma) L_m(\tilde{\psi}_1(X_{\sigma(1)},X_{\sigma(2)}), \\ &\qquad ...,\ \tilde{\psi}_m(X_{\sigma(D-1)},X_{\sigma(D)})), \end{aligned} \tag{3.2}$$

for $X_1,...,X_D \in T_u(P)$ (tangent space of P at u), where the summation is taken over all permutations σ of $(1,...,D)$ and $\epsilon(\sigma)$ is the sign of the permutation. If $\{T_A\}$ is a basis for $\mathcal{G}$ such that $\tilde{\psi}_i = \tilde{\psi}_i^A T_A$ [$\tilde{\psi}_i^A \in \Lambda^2(P,\mathbb{R})$ the space of real-valued two-forms on P], then one obtains from the multilinearity of L_m and the definition of the wedge product that

$$L_m(\tilde{\psi}_1,...,\tilde{\psi}_m) = L_m(T_{A_1},...,T_{A_m}) \tilde{\psi}_1^{A_1} \wedge \cdots \wedge \tilde{\psi}_m^{A_m}. \tag{3.3}$$

If the restrictions $\psi_i = \gamma^* \tilde{\psi}_i$ are tensorial forms on Q, then the restriction of the Weil form itself, i.e.,

$$L_m(\psi_1,...,\psi_m) = \gamma^* L_m(\tilde{\psi}_1,...,\tilde{\psi}_m) \tag{3.4}$$

will project to a unique D-form $\mathcal{L}$ on M such that (Ref. 4, Chap. XII)

$$L_m(\psi_1,...,\psi_m) = \pi^* \mathcal{L}, \tag{3.5}$$

where π is the projection from the subbundle Q on the base manifold M.

To determine the dynamics of a gauge theory, a gauge field Lagrangian must be defined. The gauge invariant Lagrangian form on $P(M,G)$, M the space-time manifold, given by

$$L_2(\tilde{\Delta},*\tilde{\Delta}), \tag{3.6}$$

where $\tilde{\Delta}$ is the curvature two-form given in Eq. (2.5), $*$ the Hodge duality transformation, and L_2 the Weil polynomial formed by summation over algebraically independent elements of $S^2_G(\mathcal{G})$, will lead to the electroweak gauge field Lagrangian if $G=\mathrm{U}(2)$, and to a gravitational gauge field Lagrangian if $G=\mathrm{SO}(2,3)$.

The gauge invariance of (3.6), that is $L_2[\tilde{\Delta}(s^*(\tilde{\mu})),*\tilde{\Delta}(s^*(\tilde{\mu}))]=L_2(\tilde{\Delta}(\tilde{\mu}),*\tilde{\Delta}(\tilde{\mu}))$ for each diffeomorphism $s:P\to P$ such that $s(ug)=s(u)g$ and $\pi(u)=\pi(s(u))$ for each $g\in G$ and $u\in P$ (Ref. 13), follows from the $\mathrm{Ad}(G)$-invariance of L_2 and the fact that $\tilde{\Delta}(\tilde{\mu})$ is a tensorial two-form on P.[3,5]

In both cases, unitary and anti-de Sitter, the condition $[\gamma_*\mathcal{H},\mathcal{C}]\subset\mathcal{C}$ is satisfied as we will see in the next sections. Then Δ, Ω, Φ, and Σ as given in Eqs. (2.6)–(2.9) are all tensorial two-forms on Q, i.e.,

$$L_2(\Delta,*\Delta)=\gamma^*L_2(\tilde{\Delta},*\tilde{\Delta}) \tag{3.7}$$

projects to a unique four-form $\mathcal{L}$ on M such that

$$L_2(\Delta,*\Delta)=\pi^*\mathcal{L}. \tag{3.8}$$

To determine Weil polynomials on an arbitrary Lie algebra $\mathcal{G}$, use can be made of the isomorphism between the algebra $S_G(\mathcal{G})$ of symmetric $\mathrm{Ad}(G)$-invariant multilinear mappings on $\mathcal{G}$ and the algebra $P_G(\mathcal{G})$ of homogeneous $\mathrm{Ad}(G)$-invariant polynomial functions on $\mathcal{G}$. The identifying isomorphism $\mathcal{S}:P_G(\mathcal{G})\to S_G(\mathcal{G})$ is determined by[14]

$$(\mathcal{S}f)(X_1,\ldots,X_k)=a_{i_1\cdots i_k}\xi^{i_1}(X_1)\cdot\cdots\cdot\xi^{i_k}(X_k),\quad X_1,\ldots,X_k\in\mathcal{G}, \tag{3.9}$$

where $a_{i_1\cdots i_k}\xi^{i_1}\cdot\cdots\cdot\xi^{i_k}$ is the unique expression for the polynomial function of degree k in the basis $\{\xi^i\}$ for $\mathcal{G}^*$ the dual space of $\mathcal{G}$.

Algebraically independent and generating $\mathrm{Ad}(U(n))$-invariant polynomial functions on the Lie algebra $\mathrm{u}(n)$ are given by the characteristic coefficients $f_k(X)$ in[4]

$$\det(\lambda I_n+iX)=\sum_{k=0}^{n}(-1)^k f_k(X)\lambda^{n-k},\quad \text{for } X\in u(n). \tag{3.10}$$

If $n=2$ and $\{\chi^i\}$ is a basis for $\mathrm{u}(2)^*$ such that $\chi^i{}_j(X)=X^i{}_j$ for $X\in\mathrm{u}(2)$, then the polynomial function of degree 2 is given by

$$f_2(X)=-\tfrac{1}{2}(X^i{}_iX^j{}_j-X^i{}_jX^j{}_i). \tag{3.11}$$

The Weil polynomial corresponding to f_2 under the isomorphism $\mathcal{S}$ given in Eq. (3.9) is then

$$L_2(X_1,X_2)=-\tfrac{1}{2}(X^i_{1i}X^j_{2j}-X^i_{1j}X^j_{2i}),\quad \text{for } X_1,X_2\in\mathrm{u}(2). \tag{3.12}$$

So, all $\mathrm{Ad}(U(2))$-invariant Weil polynomials on $\mathrm{u}(2)$ of degree 2 are polynomials proportional to the foregoing. In the basis

$$\{T_1,T_2,T_3,T_4\}=\{\tfrac{1}{2}i\tau_1,\tfrac{1}{2}i\tau_2,\tfrac{1}{2}i\tau_3,\tfrac{1}{2}iI\}, \tag{3.13}$$

for $\mathrm{u}(2)$, where τ_1, τ_2, τ_3 are the Pauli matrices and I the 2×2 identity matrix, we find that these Weil polynomials are determined by

$$L_2(T_a,T_b)=C_1\delta_{ab},\quad a=1,2,3, \tag{3.14a}$$

$$L_2(T_4,T_4)=C_2, \tag{3.14b}$$

where we introduced constants C_1,C_2 since the Weil polynomials are defined only upon an arbitrary constant.

For $G=\mathrm{SO}(2,3)$ one can use the same methods to find that all $\mathrm{Ad}(\mathrm{SO}(2,3))$-invariant Weil polynomials on $\mathrm{so}(2,3)$ of degree 2 are polynomials proportional to

$$L_2(X_1,X_2)=\tfrac{1}{2}(X^A_{1A}X^B_{2B}-X^A_{1B}X^B_{2A}),\quad \text{for } X_1,X_2\in\mathrm{so}(2,3), \tag{3.15}$$

with $X^A_B=\chi^A_B(X)$, $\{\chi^A_B\}$ a basis for $\mathrm{so}(2,3)^*$. The generators J_{AB} of the Lie algebra $\mathrm{so}(2,3)$ are in 5×5-matrix representation given by

$$(J_{AB})^K{}_L=\delta^K{}_A\eta_{BL}-\delta^K{}_B\eta_{AL},\quad A,B,K,L=0,1,2,3,4, \tag{3.16}$$

where $\eta_{AB}=\mathrm{diag}(-1,1,1,1,-1)$. The ten elements $J_{AB}=-J_{BA}$ can be split into six generators J_{ab}, $a,b=0,1,2,3$ of the $\mathrm{SO}(1,3)$ subgroup and four anti-de Sitter boosts $P_a=(1/l)J_{4a}$, l the de Sitter length. In the basis $\{J_{ab},P_c\}$ thus defined, we obtain from (3.15) and (3.16) that

$$L_2(J_{ab},J_{cd})=C_1(\eta_{ac}\eta_{bd}-\eta_{ad}\eta_{bc})=C_1\eta_{ab,cd}, \tag{3.17a}$$

$$L_2(P_a,P_b)=C_2(1/l)^2\eta_{ab}, \tag{3.17b}$$

$$L_2(J_{ab},P_c)=0, \tag{3.17c}$$

where C_1 and C_2 are arbitrary constants.

IV. THE ELECTROWEAK GAUGE FIELD LAGRANGIAN

In the electroweak theory, the photon field and massive vector boson fields are constructed from the $\mathrm{SU}(2)\times\mathrm{U}(1)$ gauge fields and a suitable scalar Higgs field that reduces the original symmetry to a $\mathrm{U}(1)$ subsymmetry. The Lie algebra $\mathrm{u}(2)=\mathrm{su}(2)\oplus\mathrm{u}(1)$ of $\mathrm{SU}(2)\times\mathrm{U}(1)$ is, however, also the Lie algebra of $\mathrm{U}(2)$. A change from $\mathrm{SU}(2)\times\mathrm{U}(1)$ to $\mathrm{U}(2)$ symmetry will therefore alter only few of the detailed mass calculations in the model (see remark in Ref. 11). Throughout the remainder of the section we then assume that (i) $P(M,G)$ is a $\mathrm{U}(2)$ principal fiber bundle over four-dimensional space-time M; (ii) $Q(M,H)$ is a $\mathrm{U}(1)$ subbundle of P. More precisely, the subgroup H is the charge conservation subgroup $\mathrm{U}_c(1)$ of $\mathrm{U}(2)$, i.e.,

$$A\in H \text{ if and only if } A=\begin{pmatrix}e^{2i\theta}&0\\0&1\end{pmatrix}. \tag{4.1}$$

The structure group $\mathrm{U}_c(1)$ of Q is the isotropy subgroup of $\mathrm{U}(2)$ at points $\binom{0}{a}\in\mathbb{C}^2$, $a\in\mathbb{R}^+$ (Ref. 11). The orbit space $V_0=\mathrm{U}(2)\cdot\binom{0}{a}$ is $\bar{S}$, the sphere of radius a in $\mathbb{C}^2$, and is diffeomorphic with the coset manifold $\mathrm{U}(2)/\mathrm{U}_c(1)$. Therefore, the existence of the reduced subbundle Q of P implies the existence of a global section in the vector bundle $E(M,\bar{S},\mathrm{U}(2),P)$ associated with P, or equivalently of a symmetry breaking Higgs field $\psi:P\to\bar{S}$ such that $\psi^{-1}\binom{0}{a}=Q$.

In the basis (3.13), a connection one-form $\tilde{\mu}$ on P may be written as

$$\tilde{\mu} = \tilde{a}^1 g T_1 + \tilde{a}^2 g T_2 + \tilde{a}^3 g T_3 + \tilde{a}^4 g' T_4, \tag{4.2}$$

where g and g' are coupling constants. We then define the new basis $\{\overline{T}_1, \overline{T}_2, \overline{T}_3, \overline{T}_4\}$, where

$$\overline{T}_1 = (gT_1 - igT_2)/\sqrt{2}, \tag{4.3}$$

$$\overline{T}_2 = (gT_1 + igT_2)/\sqrt{2}, \tag{4.4}$$

$$\overline{T}_3 = -\cos\alpha(gT_3) + \sin\alpha(g'T_4), \tag{4.5}$$

$$\overline{T}_4 = \sin\alpha(gT_3) + \cos\alpha(g'T_4) \tag{4.6}$$

In this basis, the restriction μ of $\tilde{\mu}$ to Q is written as

$$\mu = W^- \overline{T}_1 + W^+ \overline{T}_2 + Z\overline{T}_3 + A\overline{T}_4, \tag{4.7}$$

where

$$W^- = (a^1 + ia^2)/\sqrt{2}, \tag{4.8}$$

$$W^+ = (a^1 - ia^2)/\sqrt{2}, \tag{4.9}$$

$$Z = -\cos\alpha\, a^3 + \sin\alpha\, a^4, \tag{4.10}$$

$$A = \sin\alpha\, a^3 + \cos\alpha\, a^4. \tag{4.11}$$

It is clear that we identify α as the Weinberg angle θ_w, such that the charge e of the electromagnetic field A is given by

$$e = g\sin\alpha = g'\cos\alpha = gg'[g^2 + (g')^2]^{-1/2}. \tag{4.12}$$

Since $\overline{T}_4 = e(T_3 + T_4)$ generates the Lie algebra $\mathcal{H} = u_c(1)$ of the charge conservation subgroup, and since the nonzero commutators of the new basis satisfy

$$[\overline{T}_1, \overline{T}_4] = ie\overline{T}_1, \tag{4.13}$$

$$[\overline{T}_2, \overline{T}_4] = -ie\overline{T}_2, \tag{4.14}$$

$$[\overline{T}_1, \overline{T}_2] = ig(\cos\alpha\, \overline{T}_3 - \sin\alpha\, \overline{T}_4), \tag{4.15}$$

$$[\overline{T}_1, \overline{T}_3] = -ig\cos\alpha\, \overline{T}_1, \tag{4.16}$$

$$[\overline{T}_2, \overline{T}_3] = ig\cos\alpha\, \overline{T}_2, \tag{4.17}$$

we can, according to Sec. II, decompose μ into two pieces:

$$\mu = \omega + \phi, \tag{4.18}$$

where $\omega = A\overline{T}_4$ is a connection one-form on Q corresponding to the electromagnetic gauge field A, and $\phi = W^- \overline{T}_1 + W^+ \overline{T}_2 + Z\overline{T}_3$ is a tensorial one-form on Q whose components correspond to the vector bosons W^+, W^-, and Z (see also Ref. 11).

The reduction to Q of the curvature obtained from a connection one-form on P and given in Eqs. (2.6)–(2.9) is calculated making use of the commutation relations (4.13)–(4.17). We find that the components in $\Delta = \Omega + \Phi + \Sigma$ are

$$\Omega = (dA)\overline{T}_4, \tag{4.19}$$

$$\Phi = (DW^-)\overline{T}_1 + (DW^+)\overline{T}_2 + (dZ)\overline{T}_3, \tag{4.20}$$

$$\Sigma = -i\cos\theta_w(W^- \wedge Zg\overline{T}_1 - W^+ \wedge Zg\overline{T}_2 - W^- \wedge W^+ g\overline{T}_3 + W^- \wedge W^+ g'\overline{T}_4), \tag{4.21}$$

where

$$DW^- = dW^- - ieA \wedge W^-, \tag{4.22}$$

$$DW^+ = dW^+ + ieA \wedge W^+. \tag{4.23}$$

Equation (2.11) is now the Bianchi identity for the electromagnetic field and (2.12) the corresponding identity for the intermediate vector boson field.

To obtain the electroweak gauge field Lagrangian we must substitute the explicit expressions for the components Ω, Φ, Σ of Δ as given in Eqs. (4.19)–(4.23) into the Lagrangian (3.8) and apply Eq. (3.3). Then we need the expressions for $L_2(\overline{T}_i, \overline{T}_j), i,j = 1,2,3,4$, which one obtains from Eqs. (3.14) and the definitions (4.3)–(4.6) for the basis $\{\overline{T}_i\}$. To obtain dimensionless numbers for $L_2(\overline{T}_i, \overline{T}_j)$ we put $C_1 = -1/(2g^2)$ and $C_2 = -1/(2g'^2)$ in (3.14), which are also chosen such that in the final Lagrangian terms appear with the correct coefficients. Then we find that

$$L_2(\overline{T}_1, \overline{T}_2) = L_2(\overline{T}_3, \overline{T}_3) = L_2(\overline{T}_4, \overline{T}_4) = -1/2 \tag{4.24}$$

and all other zero. For the Lagrangian we obtain

$$L_2(\Delta, *\Delta) = \mathcal{L}^\gamma + \mathcal{L}^W + \mathcal{L}^Z + \mathcal{L}^{W,W} + \mathcal{L}^{W,\gamma,Z}, \tag{4.25}$$

where

$$\mathcal{L}^\gamma = -\tfrac{1}{2} dA \wedge *(dA) \tag{4.26}$$

is the Lagrangian of the electromagnetic field,

$$\mathcal{L}^W = -dW^- \wedge *(dW^+) \tag{4.27}$$

is the Lagrangian of the free intermediate charged vector boson field without the mass term,

$$\mathcal{L}^Z = -\tfrac{1}{2} dZ \wedge *(dZ) \tag{4.28}$$

is the Lagrangian of the free intermediate neutral vector field without the mass term,

$$\mathcal{L}^{W,W} = \tfrac{1}{2} g^2 W^- \wedge W^+ \wedge *(W^- \wedge W^+) \tag{4.29}$$

is the self-interaction term of the W bosons, and

$$\begin{aligned}\mathcal{L}^{W,\gamma,Z} = &-\cos\theta_w \{ i(g\, dZ - g'\, dA) \wedge *(W^- \wedge W^+) + i(gZ - g'A) \wedge [*(dW^+) \wedge W^- - *(dW^-) \wedge W^+] \\ &+ \cos\theta_w [gg'(A \wedge W^- \wedge *(W^+ \wedge Z) + Z \wedge W^- \wedge *(W^+ \wedge A)) \\ &+ g'^2 A \wedge W^- \wedge *(A \wedge W^+) + g^2 Z \wedge W^- \wedge *(Z \wedge W^+)]\}\end{aligned} \tag{4.30}$$

is the interaction between the electromagnetic, Z, and W boson fields.

In the electroweak theory, the gauge symmetry is broken spontaneously. Therefore, the gauge field Lagrangian $\mathcal{L}$ on M, of the type determined in (3.8), is supplemented with a G-invariant Lagrangian $\mathcal{L}^\Psi$ associated with a doublet of complex scalar physical Higgs fields $\Psi: M \to \mathbb{C}^2 \supset \bar{S}(a)$ with quartic self-interaction potential that

assumes a minimum for $\Psi^\dagger\Psi = a^2$ [i.e., for $\Psi(x) \in \bar{S}(a)$]. The effect of the gauge fields entering the covariant derivatives of the Higgs fields in $\mathscr{L}^\Psi$ is to give the W and Z fields a mass term such that $m_W{}^2 = a^2 g^2/2$ and $m_Z{}^2 = a^2(g^2 + g'^2)/2$ (Ref. 15).

V. GRAVITATIONAL GAUGE FIELD LAGRANGIAN

The derivation of the gravitational gauge field Lagrangian is based on the anti-de Sitter group $G = \mathrm{SO}(2,3)$ and its subgroup $H = \mathrm{SO}(1,3)$. Fundamental ingredients are (i) an anti-de Sitter principal bundle $P(M,G)$ over four-dimensional space-time M and a connection one-form $\tilde{\mu}$ on P; (ii) a Lorentz subbundle $Q(M,H)$ of P.

However, to interpret the resulting theory as a gravitational theory, the restricted curvature $\Delta = \gamma^* \tilde{\Delta}$ on Q must be directly related to the curvature of the underlying space-time manifold. Therefore, $Q(M,H)$ must be the bundle $O(M)$ of orthonormal frames over M, and $P(M,G)$ the anti-de Sitter frame bundle.

The coset space G/H is now the noncompact anti-de Sitter space $F = \mathrm{SO}(2,3)/\mathrm{SO}(1,3)$. It is a space of constant curvature $-1/l^2$ (l the curvature radius) that can be represented by a hypersurface $H_1^4(l) = \{\xi \in \mathbb{R}_2^5; \langle \xi, \xi \rangle = -l^2\}$ in $\mathbb{R}_2^5$, $\langle , \rangle$ being the quadratic form on $\mathbb{R}_2^5$ with signature $(-,+,+,+,-)$. The coset space F can also be thought as the orbit of the H-fixed point $\xi_0 = (0,0,0,0,l) \in \mathbb{R}_2^5$, and $\mathrm{SO}(1,3)$ is the isotropy subgroup of $\mathrm{SO}(2,3)$ in ξ_0. The existence of the reduced subbundle Q of P now implies the existence of a global section in the with P associated anti-de Sitter vector bundle $E(M, H_1^4, \mathrm{SO}(2,3), P)$ or of a symmetry breaking Higgs field $\psi: P \to H_1^4$ such that $\psi^{-1}(\xi_0) = Q$.

The Lie algebra $\mathscr{G} = \mathrm{so}(2,3)$ is defined by

$$[J_{ab}, J_{cd}] = J_{ad}\eta_{bc} + J_{bc}\eta_{ad} - J_{ac}\eta_{bd} - J_{bd}\eta_{ac}, \tag{5.1}$$

$$[J_{ab}, P_c] = P_a \eta_{bc} - P_b \eta_{ac}, \tag{5.2}$$

$$[P_a, P_b] = (1/l^2) J_{ab}. \tag{5.3}$$

The parameter l is still the radius of curvature of the anti-de Sitter space whose group of isomorphisms is generated by this anti-de Sitter algebra. The generators J_{ab} span the subalgebra $\mathscr{H} = \mathrm{so}(1,3)$ of $\mathscr{G}$. The P_a span a vector space $\mathscr{C} = \mathbb{R}_1^4$, such that Eq. (2.3) and the condition $[\gamma^* \mathscr{H}, \mathscr{C}] \subset \mathscr{C}$ are satisfied. Therefore, if $\tilde{\mu}$ is a connection one-form on P, its reduction μ to Q splits as in Eq. (2.4). Explicitly we may write

$$\mu = \omega + \phi = \tfrac{1}{2}\omega^{ab} J_{ab} + \phi^a P_a, \tag{5.4}$$

where now ω is a $\mathscr{H}$-valued Lorentz connection while ϕ the tensorial one-form of type $(\mathrm{Ad}\, H, \mathscr{C})$ on Q will be identified with the canonical form θ on $Q = O(M)$. Indeed, on $O(M)$ the canonical or soldering form θ has the same transformation low (2.1) as ϕ in (5.4). Therefore, it is possible to define a connection $\tilde{\mu}$ on P for which $\phi = \theta$ (Ref. 16). This identification implies that we take μ as the restriction to Q of a Cartan connection $\tilde{\mu}$ on P and that the associated anti-de Sitter vector bundle E is a soldered bundle:[17] for all $x \in M$, the fiber F_x over x is tangent to the base space M at ξ_0 and the tangent spaces $T_{\xi_0}(F_x)$ and $T_x(M)$ can be identified by an isomorphism. Moreover, the zero section $\xi(x) = \xi_0$ on E can be identified with the base space M. After identification of ϕ with the canonical form θ on Q, Ω and Φ as given in Eqs. (2.7) and (2.8) can, respectively, be identified with the curvature form Ω and torsion form Θ of the bundle $O(M)$. Explicitly, we can write

$$\Delta = \Omega + \Theta + \Sigma = \tfrac{1}{2}[\Omega^{ab} + (1/l^2)\theta^a \wedge \theta^b] J_{ab} + \Theta^c P_c, \tag{5.5}$$

where

$$\Omega^{ab} = D\omega^{ab} = d\omega^{ab} + \omega^{ac} \wedge \omega_c^b, \tag{5.6}$$

$$\Theta^a = D\theta^a = d\theta^a + \omega^{ac} \wedge \theta_c. \tag{5.7}$$

From the Bianchi identities (2.11) and (2.12) we find

$$D\Omega^{ab} = d\Omega^{ab} + \omega^{ac} \wedge \Omega_c^b - \Omega^{ac} \wedge \omega_c^b = 0, \tag{5.8}$$

$$D\Theta^a = d\Theta^a + \omega^a{}_b \wedge \Theta^b = \Omega^{ab} \wedge \theta_b. \tag{5.9}$$

Substituting the expression (5.5) for Δ into the Lagrangian form (3.8) on Q, we find making use of (3.3) and Eqs. (3.17) that

$$\begin{aligned} L_2(\Delta, {*}\Delta) = {} & (C_1/2)\Omega_{ab} \wedge {*}\Omega^{ab} \\ & + (C_1/2l^2)\epsilon_{abcd}\Omega^{ab} \wedge \theta^c \wedge \theta^d \\ & + (C_1/4l^4)\epsilon_{abcd}\theta^a \wedge \theta^b \wedge \theta^c \wedge \theta^d \\ & + (C_2/l^2)\Theta^a \wedge {*}\Theta_a, \end{aligned} \tag{5.10}$$

where $\epsilon_{0123} = 1$. Except for the coefficients, this Lagrangian is the same as the SO(1,4) gauge field Lagrangian obtained by Townsend who used the diagonal η_{ab} to contract group indices, or the Lagrangian of Zardecki who used the SO(1,4) Cartan metric to contract the Yang–Mills indices.[7] Sources for the gravitational field must be described by a matter Lagrangian $\mathscr{L}_M$. The energy-momentum three-form $T_a = T_a{}^b \epsilon_b$ and spin angular momentum three-form $S_{ab} = S_{ab}{}^c \epsilon_c$ of the matter fields, where $\epsilon_a = (1/3!)\epsilon_{abcd}\theta^b \wedge \theta^c \wedge \theta^d$, are defined through

$$\delta_\theta \mathscr{L}_M = \delta\theta^a \wedge T_a, \tag{5.11}$$

$$\delta_\omega \mathscr{L}_M = \delta\omega^{ab} \wedge S_{ab}. \tag{5.12}$$

If the total Lagrangian is varied with respect to θ^a and ω^{ab}, respectively, the following field equations are obtained:

$$\tfrac{1}{2}\epsilon_{abcd}\Omega^{ab} \wedge \theta^c - 16\pi C_2 D{*}\Theta_d = \kappa(T_d + T_d{}^{\mathrm{vac}}), \tag{5.13}$$

$$\begin{aligned} & \tfrac{1}{2}\epsilon_{abcd}\Theta^c \wedge \theta^d + 16\pi C_2 {*}\Theta_a \wedge \theta_b \\ & \quad + (\kappa/16\pi) D{*}\Omega_{ab} = -\kappa S_{ab}, \end{aligned} \tag{5.14}$$

where we have put $C_1 = 1/16\pi$, $\kappa = 8\pi l_p{}^2$, and identified l with l_p the Planck length, such that for zero torsion the field equation (5.13) reduces to Einstein's field equation. The second equation is an extension of the equation introduced by Loos[18] to describe spin density in Einstein's theory. Putting $C_2 = -1/16\pi$ would yield field equations obtainable from the SO(2,3) variant of the Zardecki action.[7] In the right-hand side of the first field equation, $T_d{}^{\mathrm{vac}}$ represents the contribution of the "θ field" to the energy momentum of the vacuum. In fact

$$T_d{}^{\mathrm{vac}} = -(1/16\pi l_p{}^4)\epsilon_{abcd}\theta^a \wedge \theta^b \wedge \theta^c, \tag{5.15}$$

which yields an enormous (negative) energy density for the vacuum. However, zero-point energies of the normal modes

of matter fields can yield comparable vacuum energy densities of the opposite sign (see for example Ref. 19), at least if general relativity is assumed to be valid upon the Planck scale. Therefore, it would be rather unnatural to obtain a small cosmological constant (see also Townsend, Ref. 7). The problem in particle physics is why all contributions to $T_d{}^{\text{vac}}$ add up to cancel exactly resulting in a zero effective cosmological constant today.[19] The fact that in (5.13) the gravitational constant $l_p{}^2$ appears in a natural way as the consequence of the commutation relations of the anti-de Sitter group has already been emphasized by Townsend.[7] Finally, we see that in vacuum and for a zero effective cosmological constant, that is $T_d{}^{\text{vac}} = 0$, the field equations (5.13) and (5.14) reduce for zero torsion to Einstein's field equations in vacuum. This is because the second equation yields in this case Yang's equation which is weaker than Einstein's equation in vacuum.[20]

If the SO(2,3) gauge theory is broken spontaneously, i.e., there exists a scalar field $\Psi: M \to \mathbb{R}_2^5 \supset H_1^4(l)$ such that the potential $V(\Psi)$ in a gauge invariant Lagrangian L^{Ψ}, associated with Ψ, assumes a minimum for $\langle \Psi, \Psi \rangle = -l^2$, that is for $\Psi(x) \in H_1^4(l)$, then the mass term given to the θ field and induced by the SO(2,3) covariant derivatives of Ψ in L^{Ψ} will give another contribution to the vacuum energy density.

VI. DISCUSSION

In the derivation of the electroweak and gravitational gauge field Lagrangian in Secs. IV and V it has become clear that fiber bundle reduction and the closely related concept of (possibly spontaneous) symmetry breaking, described in Sec. II, are essential for both theories. Moreover, this geometrical framework makes it possible to describe both theories with one type of Lagrangian, Eq. (3.6). Fields in both theories of the same geometrical origin are the W and Z vector boson fields and the soldering form θ for the tensorial component of the reduced connection, and the electromagnetic field A and Lorentz connection ω for the connection part. Other geometrical concepts that correspond are the soldering of the anti-de Sitter bundle, which is essential in the interpretation of the SO(2,3) gauge theory as a gravitation theory, and the process of symmetry breaking in electroweak theory. In the electroweak theory the manifold of vacuum states is the sphere $\bar{S}(a)$ in $\mathbb{C}^2$ while the vacuum state corresponds to the global section of $E(M, \bar{S}(a), U(2), P)$ determined by $\Psi(x) = \binom{0}{a}$. In the SO(2,3) gravitational gauge theory the role of the vacuum manifold is played by the anti-de Sitter space $H_1^4(l) \subset \mathbb{R}_2^5$ while the vacuum state corresponds to the global section of the anti-de Sitter bundle $E(M, H_1^4(l), \text{SO}(2,3), P)$ determined by $\Psi(x) = \xi_0 = (0,l)$, which section can be identified with the space-time M such that the anti-de Sitter bundle is "soldered." Then, in a spontaneously broken SO(2,3) gauge theory, four-dimensional space-time could be interpreted as the "vacuum expectation value" of a physical Higgs field $\Psi: M \to \mathbb{R}_2^5$ that breaks SO(2,3) symmetry down to the SO(1,3) subsymmetry thereby creating what is called gravitational interaction.

ACKNOWLEDGMENTS

I thank Professor D. K. Callebaut for his constant encouragement and Dr. P. Wils for reading the manuscript.

[1] A. Trautman, Rep. Math. Phys. **1**, 29 (1970).
[2] C. Nash and S. Sen, *Topology and Geometry for Physicists* (Academic, London, 1983).
[3] T. Verwimp, Prog. Theor. Phys. **80**, 330 (1988).
[4] S. Kobayashi and K. Nomizu, *Foundations of Differential Geometry* (Interscience, New York, 1963), Vols. 1 and 2.
[5] K. Kakazu and S. Matsumoto, Prog. Theor. Phys. **78**, 166, 932 (1987); **79**, 1431 (1988); **80**, 1109 (1988).
[6] D. Lovelock, J. Math. Phys. **12**, 498 (1971).
[7] P. K. Townsend, Phys. Rev. D **15**, 2795 (1977); A. Zardecki, J. Math. Phys. **29**, 1661 (1988).
[8] A. O. Barut and A. J. Bracken, ICTP Trieste, preprint IC/88/26 (1988).
[9] R. Gilmore, *Lie Groups, Lie Algebras, and Some of Their Applications* (Wiley–Interscience, New York, 1974).
[10] A. Trautman, Czech. J. Phys. B **29**, 107 (1979).
[11] R. O. Fulp and L. K. Norris, J. Math. Phys. **24**, 1871 (1983).
[12] R. Cianci, J. Math. Phys. **22**, 2759 (1981).
[13] D. Bleecker, *Gauge Theory and Variational Principles* (Addison-Wesley, Reading, MA, 1981).
[14] M. Spivak, *A Comprehensive Introduction to Differential Geometry* (Publish or Perish, Berkeley, CA, 1979), Vols. 1–5.
[15] J. Leite Lopes, *Gauge Field Theories, An Introduction* (Pergamon, New York, 1981).
[16] P. K. Smrz, J. Math. Phys. **28**, 2824 (1987).
[17] W. Drechsler, J. Math. Phys. **18**, 1358 (1977).
[18] H. G. Loos, Ann. Phys. (NY) **25**, 91 (1963).
[19] S. Weinberg, Rev. Mod. Phys. **61**, 1 (1989).
[20] E. E. Fairchild, Jr., Phys. Rev. D **14**, 384 (1976).

Conformal mapping and vertex operators in the light-cone gauge

Seichi Naito and Masahiro Nishimoto
Department of Physics, Osaka City University, Sumiyoshiku, Osaka, Japan

(Received 27 June 1989; accepted for publication 18 July 1990)

Using Fourier coefficients of Neumann functions for the tree scattering in the light-cone gauge, an open-string field operator $\Psi^p(X(x),\{\alpha_{-m}\})$ generating all "off-shell" vertex operators, which are reduced in the mass-shell case to already proposed general vertex operators, is introduced.

I. INTRODUCTION AND PRELIMINARIES

In the light-cone gauge,[1] Mandelstam[2] has derived the formula for calculating tree scattering amplitudes $A(1,2,...,M)$ among M bosonic open-string (excited) states. The result can be rewritten in the following form:[3]

$$\begin{aligned} &A(1,2,...,M) \\ &\quad = g^{M-2}G\cdot(x_M - x_1)(x_{M-1} - x_1)(x_M - x_{M-1}) \\ &\qquad \times \int dx_2\, dx_3 \cdots dx_{M-2} \prod_{r=1}^{M-1} \theta(x_{r+1} - x_r) \\ &\qquad \times \prod_{r<s} |x_r - x_s|^{p_r\cdot p_s} \\ &\qquad \times \langle E | \exp\left\{ \frac{1}{2} \sum_{r,s=1}^{M} {\sum_{m,n=0}^{\infty}}' \tilde{N}_{m,n}^{(r,s)} \alpha_{-m}^{(r)j} \alpha_{-n}^{(s)j} \right\} |0\rangle, \end{aligned} \tag{1.1}$$

where

$$p_r\cdot p_s \equiv -p_r^+ p_s^- - p_r^- p_s^+ + \sum_{j=1}^{24} p_r^j p_s^j,$$

$$p_r^\pm \equiv (1/\sqrt{2})(p_r^0 \pm p_r^{25})$$

and p_r^j $(j = 1\text{–}24)$ being the energy momentum of the rth string. In (1.1) and hereafter Σ' denotes the sum without the term with $m = n = 0$, while dummy suffix j is to be summed over $j = 1\text{–}24$. Real parameters x_r $(r = 1\text{–}M)$ are those specifying conformal mapping from upper complex z plane into the striplike ρ plane:

$$\rho(\equiv \tau + i\sigma) = \sum_{r=1}^{M} \alpha_r \ln(x_r - z), \tag{1.2}$$

where $\alpha_r \equiv 2p_r^+$, so that

$$\sum_{r=1}^{M} \alpha_r = 0. \tag{1.3}$$

Operators $\alpha_m^{(r)j}$ $(m = \pm 1, \pm 2, ...)$ satisfy

$$[\alpha_m^{(r)j}, \alpha_n^{(s)k}] = m\delta_{m+n}\delta_{rs}\delta_{jk}, \tag{1.4}$$

while $\alpha_0^{(r)j} \equiv p_r^j$.

Then, $|0\rangle$ and $\langle 0|$ are defined by

$$\alpha_n^{(r)j}|0\rangle = \langle 0|\alpha_{-n}^{(r)j} = 0 \quad (r = 1\text{–}M,\ n = 1,2,...). \tag{1.5}$$

Without loss of generality, external M bosonic open-string state $\langle E|$ can be taken to be

$$\langle E| = \langle 0| \prod_{r=1}^{M} \prod_{n=1}^{\infty} \prod_{j=1}^{24} \frac{(-i\alpha_n^{(r)j})^{e(r,n,j)}}{\sqrt{n^{e(r,n,j)}\cdot e(r,n,j)!}}, \tag{1.6}$$

$e(r,n,j)$'s being non-negative integers specifying excited state and (mass)^2 of rth string is given by

$$-p_r^2 = 2\sum_{n=1}^{\infty}\sum_{j=1}^{24} n\cdot e(r,n,j) - 2. \tag{1.7}$$

Furthermore, $\tilde{N}_{m,n}^{(r,s)}$'s in (1.1) are defined as coefficients in the following expansions;

$$\ln(z_r - x_s) = \ln(x_r - x_s) + \sum_{m=1}^{\infty} \tilde{N}_{m,0}^{(r,s)} (E^{\zeta_r})^m \quad (r \neq s), \tag{1.8}$$

$$\ln(z_r - x_r) = \zeta_r - \sum_{t(\neq r)}^{\infty} \frac{p_t^+}{p_r^+} \ln|x_r - x_t| + \sum_{m=1}^{\infty} \tilde{N}_{m,0}^{(r,r)} (E^{\zeta_r})^m, \tag{1.9}$$

$$\ln(z_r - z_s) = \ln(x_r - x_s) + {\sum_{m,n=0}^{\infty}}' \tilde{N}_{m,n}^{(r,s)} (E^{\zeta_r})^m (E^{\zeta_s})^n \quad (r \neq s) \tag{1.10}$$

and

$$\ln(z_r - \tilde{z}_r) = \ln(E^{\zeta_r} - E^{\tilde{\zeta}_r}) + {\sum_{m,n=0}^{\infty}}' \tilde{N}_{m,n}^{(r,r)} (E^{\zeta_r})^m (E^{\tilde{\zeta}_r})^n, \tag{1.11}$$

where[3]

$$E^{\zeta_r} \equiv e^{\zeta_r} \prod_{t(\neq r)} |x_t - x_r|^{-\alpha_t/\alpha_r}. \tag{1.12}$$

In (1.8)–(1.11), $z_r(\tilde{z}_r)$ represents z in the neighborhood of x_r and mapping from $z_r(\tilde{z}_r)$ into $\zeta_r(\tilde{\zeta}_r)$ is given by

$$\begin{aligned} &\alpha_r\zeta_r(\equiv \rho + i\pi(\alpha_1 + \alpha_2 + \cdots + \alpha_r)) \\ &\quad = \alpha_1 \ln(z_r - x_1) + \cdots + \alpha_r \ln(z_r - x_r) \\ &\qquad + \alpha_{r+1} \ln(x_{r+1} - z_r) + \cdots + \alpha_M \ln(x_M - z_r). \end{aligned} \tag{1.13}$$

In Sec. II and the Appendix, explicit formulas for $\tilde{N}_{m,n}^{(r,s)}$'s in (1.1) are obtained by using Lagrange's theorem. In Sec. III, we derive an open-string field operator $\Psi^p(X(x),\{\alpha_{-m}\})$ generating "off-shell" vertex operators

 0022-2488/90/123053-07$03.00

for arbitrary excited states. We also find how to obtain scattering amplitudes (1.1) by using $\Psi^p(X(x),\{\alpha_{-m}\})$. In Sec. IV, we discuss our results and suggest that our expression for $\Psi^p(X(x),\{\alpha_{-m}\})$ would be valid beyond the tree approximation. We also speculate on possibility of formulating bosonic string field theory in the complex z plane.

II. EXPLICIT FORMULAS FOR $\tilde{N}^{(r,s)}_{m,n}$

The purpose of this section is to obtain explicit formulas for $\tilde{N}^{(r,s)}_{m,n}$, with the help of which vertex operators for excited states will be explicitly constructed in Sec. III.

It has already been pointed out that $\tilde{N}^{(r,s)}_{m,n}$ in (1.1) can be calculated[4] by using Lagrange's theorem in the following special form: (1.13) gives

$$y = e^{\xi_r}\prod_{t=1}^{r-1}(x_r - x_t + y)^{-\alpha_t/\alpha_r} \times \prod_{u=r+1}^{M}(x_u - x_r - y)^{-\alpha_u/\alpha_r} \equiv e^{\xi_r}\Phi(y), \tag{2.1}$$

where $z_r \equiv x_r + y$.

Then, an arbitrary function $f(y)$ can be expanded as

$$f(y) = f(0) + \sum_{n=0}^{\infty}\frac{e^{(n+1)\xi_r}}{n+1}C_n(n+1), \tag{2.2}$$

where coefficients C's are given by

$$\sum_{j=1}^{\infty}C_j(n)y^j \equiv f'(y)\Phi^n(y) = f'(y)\prod_{t(\neq r)}|x_t - x_r|^{-n(\alpha_t/\alpha_r)} \times \exp\left(n\sum_{m=1}^{\infty}\frac{1}{m}\frac{I^+(m;x_r)}{p_r^+}y^m\right), \tag{2.3}$$

with

$$I^+(m;x_r) \equiv \sum_{s(\neq r)}\frac{p_s^+}{(x_s - x_r)^m}. \tag{2.4}$$

In the special case $f(y) = \ln(x_r - x_s + y)$, we have

$$f'(y) = \frac{1}{x_r - x_s + y} = -\sum_{l=0}^{\infty}\frac{y^l}{(x_s - x_r)^{l+1}}. \tag{2.5}$$

Therefore, (2.2), (2.3), and (2.5) lead to the following explicit formulas for $\tilde{N}^{(r,s)}_{m,0}$ $(r\neq s)$:

$$\begin{aligned}\ln(x_r - x_s + y) &= \ln(x_r - x_s) + \left(-\frac{1}{x_s - x_r}\right)E^{\xi_r} + \left(-\frac{1}{2}\frac{1}{(x_s - x_r)^2} - \frac{1}{x_s - x_r}\frac{I^+(1;x_r)}{p_r^+}\right)(E^{\xi_r})^2 \\ &\quad + \left(-\frac{1}{3}\frac{1}{(x_s - x_r)^3} - \frac{1}{(x_s - x_r)^2}\frac{I^+(1;x_r)}{p_r^+} - \frac{1}{2}\frac{1}{x_s - x_r}\frac{I^+(2;x_r)}{p_r^+} - \frac{3}{2}\frac{1}{x_s - x_r}\left(\frac{I^+(1;x_r)}{p_r^+}\right)^2\right)(E^{\xi_r})^3 + \cdots \\ &\equiv \ln(x_r - x_s) + \sum_{m=1}^{\infty}\tilde{N}^{(r,s)}_{m,0}(E^{\xi_r})^m.\end{aligned} \tag{2.6}$$

Furthermore, we can see from (1.8), (1.9), and (1.13) that $\tilde{N}^{(r,r)}_{m,0}$ can be calculated by

$$\tilde{N}^{(r,r)}_{m,0} = -\frac{1}{p_r^+}\sum_{s(\neq r)}p_s^+\tilde{N}^{(r,s)}_{m,0}, \tag{2.7}$$

so that we have

$$\begin{aligned}\sum_{m=1}^{\infty}\sum_{s=1}^{M}\tilde{N}^{(r,s)}_{m,0}p_s^j\alpha^{(r)j}_{-m} &= \left(-I^j(1;x_r) + p_r^j\frac{I^+(1;x_r)}{p_r^+}\right)\alpha^{(r)j}_{-1} \\ &\quad + \left(-\frac{1}{2}I^j(2;x_r) - I^j(1;x_r)\frac{I^+(1;x_r)}{p_r^+} + \frac{1}{2}p_r^j\frac{I^+(2;x_r)}{p_r^+} + p_r^j\left(\frac{I^+(1;x_r)}{p_r^+}\right)^2\right)\alpha^{(r)j}_{-2} \\ &\quad + \left(-\frac{1}{3}I^j(3;x_r) - I^j(2;x_r)\frac{I^+(1;x_r)}{p_r^+} - \frac{1}{2}I^j(1;x_r)\frac{I^+(2;x_r)}{p_r^+} - \frac{3}{2}I^j(1;x_r)\left(\frac{I^+(1;x_r)}{p_r^+}\right)^2\right. \\ &\quad \left. + \frac{1}{3}p_r^j\frac{I^+(3;x_r)}{p_r^+} + \frac{3}{2}p_r^j\frac{I^+(2;x_r)}{p_r^+}\frac{I^+(1;x_r)}{p_r^+} + \frac{3}{2}p_r^j\left(\frac{I^+(1;x_r)}{p_r^+}\right)^3\right)\alpha^{(r)j}_{-3} + \cdots \\ &\equiv \sum_{m=1}^{\infty}\tilde{N}^{(r)j}_{m}\alpha^{(r)j}_{-m},\end{aligned} \tag{2.8}$$

where

$$I^j(m;x_r)\equiv\sum_{t(\neq r)}\frac{p_t^j}{(x_t-x_r)^m}. \tag{2.9}$$

On the other hand, $\tilde{N}_{m,n}^{(r,s)}$ $(m,n=1,2,...)$ can be obtained from $\tilde{N}_{m,0}^{(r,s)}$ [given explicitly by (2.6)] in the following way. First, (2.13) gives[4]

$$\left(\alpha_s\frac{\partial}{\partial\xi_r}+\alpha_r\frac{\partial}{\partial\xi_s}\right)\ln(z_r-z_s)=\sum_{t=1}^{M}\alpha_t\frac{\partial}{\partial\xi_r}\ln(z_r-x_t)\frac{\partial}{\partial\xi_r}\ln(z_s-x_t), \tag{2.10}$$

which together with (1.3), (1.10), (1.11), and (2.7) leads to

$$\tilde{N}_{m,n}^{(r,s)}=\frac{mn}{m\alpha_s+n\alpha_r}\sum_{t(\neq r,s)}\alpha_t(\tilde{N}_{m,0}^{(r,t)}-\tilde{N}_{m,0}^{(r,s)})\times(\tilde{N}_{n,0}^{(s,t)}-\tilde{N}_{n,0}^{(s,r)})\quad(\text{for } r\neq s) \tag{2.11}$$

and

$$\tilde{N}_{m,n}^{(r,r)}=\frac{mn}{m+n}\left(\sum_{t(\neq r)}\frac{p_t^+}{p_r^+}\tilde{N}_{m,0}^{(r,t)}\tilde{N}_{n,0}^{(r,t)}+\tilde{N}_{m,0}^{(r,r)}\tilde{N}_{n,0}^{(r,r)}\right). \tag{2.12}$$

Formulas (2.12) and (2.6) give

$$\tilde{N}_{1,1}^{(r,r)}=\frac{1}{2}\frac{I^+(2;x_r)}{p_r^+}+\frac{1}{2}\left(\frac{I^+(1;x_r)}{p_r^+}\right)^2, \tag{2.13}$$

$$\tilde{N}_{2,1}^{(r,r)}=\frac{1}{3}\frac{I^+(3;x_r)}{p_r^+}+\frac{I^+(2;x_r)}{p_r^+}\frac{I^+(1;x_r)}{p_r^+}+\frac{2}{3}\left(\frac{I^+(1;x_r)}{p_r^+}\right)^3, \tag{2.14}$$

$$\tilde{N}_{3,1}^{(r,r)}=\frac{1}{4}\frac{I^+(4;x_r)}{p_r^+}+\frac{I^+(3;x_r)}{p_r^+}\frac{I^+(1;x_r)}{p_r^+}+\frac{3}{8}\left(\frac{I^+(2;x_r)}{p_r^+}\right)^2+\frac{9}{4}\frac{I^+(2;x_r)}{p_r^+}\times\left(\frac{I^+(1;x_r)}{p_r^+}\right)^2+\frac{9}{8}\left(\frac{I^+(1;x_r)}{p_r^+}\right)^4, \tag{2.15}$$

and

$$\tilde{N}_{2,2}^{(r,r)}=\frac{1}{4}\frac{I^+(4;x_r)}{p_r^+}+\frac{I^+(3;x_r)}{p_r^+}\frac{I^+(1;x_r)}{p_r^+}+\frac{1}{4}\left(\frac{I^+(2;x_r)}{p_r^+}\right)^2+2\frac{I^+(2;x_r)}{p_r^+}\left(\frac{I^+(1;x_r)}{p_r^+}\right)^2+\left(\frac{I^+(1;x_r)}{p_r^+}\right)^4, \tag{2.16}$$

etc.

As for $\tilde{N}_{m,n}^{(r,s)}$ $(r\neq s,\ m,n=1,2,...)$, we derive a formula (A5) in the Appendix, with the help of which we can easily obtain (A6)–(A9). The formula (A5) together with the observation that any $\tilde{N}_{m,n}^{(r,s)}$ (or $\tilde{N}_m^{(r)j}$) can be expressed as a polynomial of $(x_r-x_s)^{-p}$ [or $I^j(p;x_r)$] and $I^+(p;x_{r(s)})$ $(p=1\text{–}m)$ is the new contribution of this paper.

III. VERTEX OPERATORS

In the covariant string theory, we have free fields $X^\mu(x_r)$, two point function of which is given by[5]

$$\langle X^\mu(x_r)X^\nu(x_s)\rangle=-\eta^{\mu\nu}\ln(x_r-x_s). \tag{3.1}$$

Then, we have the following operator product expansions:

$$\frac{-i}{(m-1)!}\left(\frac{\partial}{\partial x_r}\right)^m X^\mu(x_r):\prod_{s(\neq r)}e^{ip_sX(x_s)}:=\left(\sum_{s(\neq r)}\frac{p_s^\mu}{(x_s-x_r)^m}\right):\prod_{s(\neq r)}e^{ip_sX(x_s)}:-\frac{i}{(m-1)!}:\left(\frac{\partial}{\partial x_r}\right)^m X^\mu(x_r)\prod_{s(\neq r)}e^{ip_sX(x_s)}:, \tag{3.2}$$

where : : represents normal ordering.

Suggested by (3.2) we introduce an open-string field operator

$$\Psi^p(X(x);\{\alpha_{-m}\})\equiv:e^{ip_\mu X^\mu(x)}\exp\left(\sum_{m=1}^{\infty}Y_m^j\alpha_{-m}^j+\frac{1}{2}\sum_{m,n=1}^{\infty}Y_{mn}\alpha_{-m}^j\alpha_{-n}^j\right):, \tag{3.3}$$

where we construct Y_m^j and Y_{mn} as polynomials of $X(x)$ by making the following replacements in $\tilde{N}_m^{(r)j}$ and $\tilde{N}_{m,n}^{(r,r)}$, respectively;

$$p_r^\mu\to p^\mu, \tag{3.4}$$

$$I^\mu(m;x_r)\to\frac{-i}{(m-1)!}\partial^m X^\mu(x). \tag{3.5}$$

For example, we find from (2.8) and (2.13)–(2.16)

$$\sum_{m=1}^{\infty}Y_m^j\alpha_{-m}^j+\frac{1}{2}\sum_{m,n=1}^{\infty}Y_{mn}\alpha_{-m}^j\alpha_{-n}=\left(\partial X^j-p^j\frac{\partial X^+}{p^+}\right)(i\alpha_{-1}^j)+\left(\frac{1}{2}\partial^2X^j-i\partial X^j\frac{\partial X^+}{p^+}-\frac{1}{2}p^j\frac{\partial^2X^+}{p^+}+ip^j\left(\frac{\partial X^+}{p^+}\right)^2\right)(i\alpha_{-2}^j)$$

$$+\left(\frac{1}{6}\partial^3X^j - i\partial^2X^j\frac{\partial X^+}{p^+} - \frac{i}{2}\partial X^j\frac{\partial^2X^+}{p^+} - \frac{3}{2}\partial X^j\left(\frac{\partial X^+}{p^+}\right)^2 - \frac{1}{6}p^j\frac{\partial^3X^+}{p^+} + \frac{3}{2}ip^j\frac{\partial^2X^+}{p^+}\frac{\partial X^+}{p^+}\right.$$
$$\left.+\frac{3}{2}p^j\left(\frac{\partial X^+}{p^+}\right)^3\right)(i\alpha^j_{-3}) + \left(\frac{i}{4}\frac{\partial^2X^+}{p^+} + \frac{1}{4}\left(\frac{\partial X^+}{p^+}\right)^2\right)(i\alpha^j_{-1})(i\alpha^j_{-1})$$
$$+\left(\frac{i}{6}\frac{\partial^3X^+}{p^+} + \frac{\partial^2X^+}{p^+}\frac{\partial X^+}{p^+} - \frac{2}{3}i\left(\frac{\partial X^+}{p^+}\right)^3\right)(i\alpha^j_{-1})(i\alpha^j_{-2}) + \cdots. \tag{3.6}$$

Substituting (3.6) into (3.3), we find

$$\Psi^p(X(x);\{\alpha_{-m}\}) = :e^{ipX(x)}: + :\left(\partial X^j - p^j\frac{\partial X^+}{p^+}\right)e^{ipX(x)}:(i\alpha^j_{-1}) + \frac{1}{2}:\left[\left(\partial X^j - p^j\frac{\partial X^+}{p^+}\right)\left(\partial X^k - p^k\frac{\partial X^+}{p^+}\right)\right.$$
$$\left.+\delta^{jk}\left(\frac{i}{2}\frac{\partial^2X^+}{p^+} + \frac{1}{2}\left(\frac{\partial X^+}{p^+}\right)^2\right)\right]e^{ipX(x)}:(i\alpha^j_{-1})(i\alpha^k_{-1})$$
$$+\frac{1}{2}:\left(\partial^2X^j - 2i\,\partial X^j\frac{\partial X^+}{p^+} - p^j\frac{\partial^2X^+}{p^+} + 2ip^j\left(\frac{\partial X^+}{p^+}\right)^2\right)e^{ipX(x)}:(i\alpha^j_{-2})$$
$$+\frac{1}{6}:\left\{\left(\partial X^j - p^j\frac{\partial X^+}{p^+}\right)\left(\partial X^k - p^k\frac{\partial X^+}{p^+}\right)\left(\partial X^l - p^l\frac{\partial X^+}{p^+}\right)\right.$$
$$+\left[\left(\partial X^j - p^j\frac{\partial X^+}{p^+}\right)\delta^{kl}\left(\frac{i}{2}\frac{\partial^2X^+}{p^+} + \frac{1}{2}\left(\frac{\partial X^+}{p^+}\right)^2\right)\right.$$
$$\left.\left.\left.+(\text{other two terms obtained by cyclic permutations among } (j,k,l))\right]\right\}e^{ipX(x)}:$$
$$\times(i\alpha^j_{-1})(i\alpha^k_{-1})(i\alpha^l_{-1})$$
$$+\frac{1}{2}:\left[\left(\partial X^j - p^j\frac{\partial X^+}{p^+}\right)\left(\partial^2X^k - 2i\partial X^j\frac{\partial X^+}{p^+} - p^j\frac{\partial^2X^+}{p^+} + 2ip^j\left(\frac{\partial X^+}{p^+}\right)^2\right)\right.$$
$$\left.+\delta^{jk}\left(\frac{i}{3}\frac{\partial^3X^+}{p^+} + 2\frac{\partial^2X^+}{p^+}\frac{\partial X^+}{p^+} - \frac{4}{3}i\left(\frac{\partial X^+}{p^+}\right)^3\right)\right]e^{ipX(x)}:(i\alpha^j_{-1})(i\alpha^k_{-2})$$
$$+\frac{1}{3}:\left(\frac{1}{2}\partial^3X^j - 3i\partial^2X^j\frac{\partial X^+}{p^+} - \frac{3}{2}i\partial X^j\frac{\partial^2X^+}{p^+} - \frac{9}{2}\partial X^j\left(\frac{\partial X^+}{p^+}\right)^2 - \frac{1}{2}p^j\frac{\partial^3X^+}{p^+}\right.$$
$$\left.+\frac{9}{2}ip^j\frac{\partial^2X^+}{p^+}\frac{\partial X^+}{p^+} + \frac{9}{2}p^j\left(\frac{\partial X^+}{p^+}\right)^3\right)e^{ipX(x)}:(i\alpha^j_{-3}) + \cdots \tag{3.7}$$
$$\equiv\sum\left(\prod_{n,j}\frac{(i\alpha^j_{-n})^{e(n,j)}}{\sqrt{n^{e(n,j)}\cdot e(n,j)!}}\right)V^p_{\{e(n,j)\}}(x), \tag{3.8}$$

where the sum in (3.8) is taken over all non-negative $e(n,j)$. Also, $V^p_{\{e(n,j)\}}(x)$ defined by (3.8) are found to be vertex operators, since (3.1)–(3.8) and (A5) lead to the following identity valid even in off-mass-shell case:

$$\left(\prod_{r<s}|x_r - x_s|^{p_r\cdot p_s}\right)\exp\left(\frac{1}{2}\sum_{r,s}{}'\sum_{m,n}\tilde{N}^{(r,s)}_{m,n}\alpha^{(r)j}_{-m}\alpha^{(s)j}_{-n}\right) = \left\langle\prod_{r=1}^{M}\Psi^{p_r}(X(x_r);\{\alpha^{(r)}_{-m}\})\right\rangle. \tag{3.9}$$

The right-hand side of (3.9) means that tree scattering amplitudes given by Mandelstam in the light-cone gauge have noncovariant features only in the polarization tensors for external excited string states. Moreover, (3.7) and (3.8) show that an open-string field operator $\Psi^p(X(x),\{\alpha_{-m}\})$ (3.3) with (3.6) generates an "off-shell" vertex operator $V^p_{\{e(n,j)\}}$ having conformal weight

$$\frac{p^2}{2} + \sum_{n=1}^{\infty}\sum_{j=1}^{24}n\cdot e(n,j)$$

[which is equal to one only under physical condition (1.7)], since we find the following operator product expansion:

$$:-\frac{1}{2}\partial_zX^\mu(z)\partial_zX_\mu(z):\Psi^p(X(x);\{\alpha_{-m}\})$$
$$=\frac{(p^2/2)\Psi^p(X(x);\{\alpha_{-m}\}) + [N,\Psi^p(X(x);\{\alpha_{-m}\})]}{(z-x)^2} + \frac{1}{z-x}\partial_x\Psi^p(X(x);\{\alpha_{-m}\}) + \cdots, \tag{3.10}$$

where N is the number operator

$$N \equiv \sum_{n=1}^{\infty} \alpha^{j}_{-n} \alpha^{j}_{n}. \tag{3.11}$$

On the other hand, "on-shell" vertex operators have already been proposed, in the following simple way:[6]

$$:e^{ip_0 \mu X^{\mu}(x)}: :A^{j}_{m}(x) :A^{k}_{n}(x): \cdots, \tag{3.12}$$

where

$$A^{j}_{m}(x) \equiv \oint_{x} \frac{dz}{2\pi i} \partial_z X^{j}(z) e^{imk_{\mu} X^{\mu}(x)} \tag{3.13}$$

with the restrictions

$$p_0{}^2 = 2, \quad k \cdot p_0 = -1, \quad k^2 = 0. \tag{3.14}$$

Applying operator product expansion to (3.12), we can explicitly convert (3.12) into the form that has only one overall normal ordering. Thus obtained results are found to be proportional to those vertex operators [multiplied by $\alpha^{j}_{-m} \alpha^{k}_{-n} \cdots$ in (3.7)], when we set

$$p = p_0 + (m + n + \cdots)k$$

in the special frame

$$k^{+} = k^{j} = 0, \quad k^{-} = 1/p_0^{+}. \tag{3.15}$$

Finally, we can explicitly calculate (3.9) for $M = 2,3$; for $M = 2$, we find from (1.10) and (3.9)

$$\begin{aligned} &\langle \Psi^{p}(X(x_2);\{\alpha^{(2)}_{-m}\}) \Psi^{-p}(X(x_1);\{\alpha^{(1)}_{-m}\}) \rangle \\ &\quad = |x_2 - x_1|^{-p^2} \exp\left(- \sum_{m=1}^{\infty} \frac{(-1)^m}{m(x_2 - x_1)^{2m}} \right. \\ &\qquad \left. \times \alpha^{(1)j}_{-m} \alpha^{(2)j}_{-m} \right), \end{aligned} \tag{3.16}$$

so that (3.8) gives

$$\begin{aligned} &\langle 0 | V^{p}_{\{e(2,m,j)\}}(x_2) V^{-p}_{\{e(1,n,k)\}}(x_1) | 0 \rangle \\ &\quad = \frac{(-1)^{N_2} |x_2 - x_1|^{-p^2 - 2N_2 + 2}}{(x_2 - x_1)^2} \\ &\qquad \times \prod_{m=1}^{\infty} \prod_{j=1}^{24} \delta_{e(2,m,j), e(1,m,j)}, \end{aligned} \tag{3.17}$$

where, and hereafter, N_r is an eigenvalue of (3.11) for the rth string:

$$N_r \equiv \sum_{m=1}^{\infty} \sum_{j=1}^{24} m \cdot e(r,m,j). \tag{3.18}$$

On the other hand, (1.13), (1.18), and (1.9) for $M = 3$ give

$$\begin{aligned} \sum_{s=1}^{3} \tilde{N}^{(1,s)}_{m,0} p^{j}_{s} &= (\alpha_1 p^{j}_{2} - \alpha_2 p^{j}_{1}) \frac{1}{\alpha_1} (\tilde{N}^{(1,2)}_{m,0} - \tilde{N}^{(1,3)}_{m,0}), \\ &= (\alpha_1 p^{j}_{2} - \alpha_2 p^{j}_{1}) \tilde{N}^{1}_{m}, \end{aligned} \tag{3.19}$$

where

$$\begin{aligned} &\alpha_1 \left(\frac{(x_2 - x_1)(x_3 - x_1)}{x_2 - x_3} \right)^{m} \tilde{N}^{1}_{m} \\ &\quad = \frac{1}{m!} \left(-m \frac{\alpha_2}{\alpha_1} - 1 \right) \left(-m \frac{\alpha_2}{\alpha_1} - 2 \right) \cdots \\ &\qquad \times \left(-m \frac{\alpha_2}{\alpha_1} - m + 1 \right) \\ &\quad \equiv \alpha_1 N^{1}_{m}. \end{aligned} \tag{3.20}$$

It is easy to see that (3.19) and (3.20) are valid even if we make cyclic permutations $1 \to 2 \to 3 \to 1$. Therefore, (2.11) and (2.12) give

$$\tilde{N}^{(r,s)}_{m,n} = \frac{-mn\alpha_1 \alpha_2 \alpha_3}{m\alpha_s + n\alpha_r} \tilde{N}^{r}_{m} \tilde{N}^{s}_{n}. \tag{3.21}$$

In conclusion, we find

$$\begin{aligned} &|x_1 - x_2|^{p_1 \cdot p_2} |x_2 - x_3|^{p_2 \cdot p_3} |x_3 - x_1|^{p_3 \cdot p_1} \left\langle E \left| \exp\left(\frac{1}{2} \sum_{r,s} \sum_{m,n}{}' \tilde{N}^{(r,s)}_{m,n} \alpha^{(r)j}_{-m} \alpha^{(s)j}_{-m} \right) \right| 0 \right\rangle \\ &\quad = \frac{(-1)^{N_1 + N_2 + N_3}}{|(x_1 - x_2)(x_2 - x_3)(x_3 - x_1)|} \left| \frac{(x_1 - x_2)(x_1 - x_3)}{x_2 - x_3} \right|^{-p_1^2/2 + N_1 - 1} \\ &\qquad \times \left| \frac{(x_2 - x_3)(x_2 - x_1)}{x_3 - x_1} \right|^{-p_2^2/2 + N_2 - 1} \left| \frac{(x_3 - x_1)(x_3 - x_2)}{x_1 - x_2} \right|^{-p_3^2/2 + N_3 - 1} \\ &\qquad \times \left\langle E \left| \exp\left(\frac{1}{2} \sum_{r,s=1}^{3} \sum_{m,n=1}^{\infty} N^{rs}_{mn} \alpha^{(r)j}_{-m} \alpha^{(s)j}_{-n} + (\alpha_1 p^{j}_{2} - \alpha_2 p^{j}_{1}) \sum_{r=1}^{3} N^{r}_{m} \alpha^{(r)j}_{-m} \right) \right| 0 \right\rangle, \end{aligned} \tag{3.22}$$

where N^{r}_{m} is defined by (3.20) and

$$N^{rs}_{mn} \equiv -\frac{mn\alpha_1 \alpha_2 \alpha_3}{m\alpha_s + n\alpha_r} N^{r}_{m} N^{s}_{n}. \tag{3.23}$$

IV. CONCLUSIONS AND DISCUSSIONS

The results we have obtained in this paper are summarized and interpreted as follows: Scattering amplitudes $A(1,2,\ldots,M)$ among M bosonic open-string (excited) states are expressed by

$$\begin{aligned} &A(1,2,\ldots,M) \\ &\quad = g^{M-2} G(x_{M-1} - x_1)\ (x_{M-1} - x_1)\ (x_M - x_{M-1}) \\ &\qquad \times \int dx_2\ dx_3 \cdots dx_{M-2} \prod_{r=1}^{M-1} \theta(x_{r+1} - x_r) \end{aligned}$$

$$\times\left\langle E\left|\left(\left\langle\prod_{r=1}^{M}\psi^{p_r}(X(x_r);\{\alpha^{(r)}_{-m}\})\right\rangle_{\text{vac}}\right)\right|0\right\rangle, \quad (4.1)$$

with

$$\psi^p(X(x);\{\alpha_{-M}\}) \equiv \sum\left(\prod_{nj}\frac{(i\alpha^j_{-n})^{e(n,j)}}{\sqrt{n^{e(nj)}\cdot e(n_j)!}}\right)V^p_{\{e(n,j)\}}(x), \quad (4.2)$$

where $V^p_{\{e(n,j)\}}(x)$ are vertex operators for excited states. It is quite remarkable that (4.2) are shown to have simple structures

$$\psi^p(X(x);\{\alpha_{-m}\}) = :e^{ip_\mu X^\mu(x)}\exp\left(\sum_{m=1}^{\infty}Y^j_m(x)\alpha^j_{-m} + \frac{1}{2}\sum_{m,n=1}^{\infty}Y_{mn}(x)\alpha^j_{-m}\alpha^j_{-n}\right):. \quad (4.3)$$

[Incidentally, expression (4.3) is possible only in the off-shell case.] Some explicit forms for Y^j_m and Y^j_{mn} are given in (3.6), so that vertex operators can be calculated as (3.7).

Although we have obtained (3.6) and (3.7) in the tree approximation, we believe that they are valid exactly (i.e., beyond tree approximation), since operator product expansions (3.10) can be derived under the assumption that

$$\langle X^\mu(x_r)X^\nu(x_s)\rangle_{\text{vac}} \sim -\eta^{\mu\nu}\ln(x_r-x_s) \quad (4.4)$$

hold in short range $x_r \sim x_s$. In (4.4), "vac" is supposed to be something like true vacuum. (In perturbation, "vac" would be expanded in loop numbers.)

At present, heterotic superstring[7] is a viable candidate for unified theory. However, since our physical space-time is four dimensional, spontaneous compactification of extra dimensions must be shown by using the nonperturbative string field theory. Before investigating this physically important problem, we speculate on bosonic string field theory formulated in z plane [i.e., not in (τ,σ) plane]. Although such a theory has not yet been constructed, open-string scattering amplitudes $A(1,2,\ldots,M)$ are calculated by (4.1) using open string field operators $\Psi^p(X(x),\{\alpha_{-m}\})$. [Together with closed string field operators $\Phi^p(\tilde{X}(z),\tilde{X}(\bar{z}),\{\tilde{\alpha}_{-m},\hat{\alpha}_{-m}\})$], $\psi^p(X(x);\{\alpha_{-m}\})$ are second quantized string field operators obeying interacting equations of motions derived from certain action principles. Furthermore, these equations should be such that they give a *kinematical* relation (4.3). Then $X^\mu(x)$ in (4.3) would be found to satisfy some *dynamical* equations of motion, with the help of which Green's functions like (4.4) would be calculated nonperturbatively. We propose that these programs would be worthy of being carried out.

APPENDIX: CALCULATION OF $\tilde{N}^{(r,s)}_{m,n}$ IN (1.1)

In much the same way as we have derived (2.6) by using (2.2) and (2.3), we obtain (for $r\neq s$)

$$\ln(z_r-z_s) = \ln(x_r-x_s-y_s) + \sum_{m=1}^{\infty}\tilde{\tilde{N}}^{(r,s)}_{m,0}(E^{\zeta_r})^m, \quad (A1)$$

where $\tilde{\tilde{N}}^{(r,s)}_{m,0}$'s are those obtained from $\tilde{N}^{(r,s)}_{m,0}$ by the following replacements:

$$\frac{1}{(x_s-x_r)^i} \to \frac{1}{(x_s+y_s-x_r)^i} \quad (i=1,2,\ldots), \quad (A2)$$

other factors $I^+(i;x_r)/p_r^+$ being unchanged. Furthermore, we find

$$\frac{1}{(x_s-x_r+y_s)^i} = \frac{-1}{(i-1)!}\left(\frac{\partial}{\partial x_r}\right)^i\ln(x_s-x_r+y_s) = \frac{-1}{(i-1)!}\left(\frac{\partial'}{\partial x_r}\right)^i\left(\ln(x_s-x_r) + \sum_{n=1}^{\infty}\tilde{N}^{(s,r)}_{n,0}(E^{\zeta_s})^n\right), \quad (A3)$$

where "primed derivatives" $(\partial'/\partial x_r)^i$ are assumed to operate only on factors $1/(x_r-x_s)^j$ in $\tilde{N}^{(s,r)}_{n,0}$ and not on any factor $I^+(j;x_s)/p_s^+$ $(j=1-n)$. In conclusion, we have from (A1)–(A3) and (2.6)

$$\begin{aligned}\ln(z_r-z_s) &= \ln(x_r-x_s-y_s) + E^{\zeta_r}\left(\frac{\partial'}{\partial x_r}\right)\ln(x_s-x_r+y_s) + (E^{\zeta_r})^2\left(\frac{1}{2}\left(\frac{\partial'}{\partial x_r}\right)^2 + \frac{I^+(1;x_r)}{p_r^+}\left(\frac{\partial'}{\partial x_r}\right)\right)\\ &\times\ln(x_s-x_r+y_s) + (E^{\zeta_r})^3\left(\frac{1}{6}\left(\frac{\partial'}{\partial x_r}\right)^3 + \frac{I^+(1;x_r)}{p_r^+}\left(\frac{\partial'}{\partial x_r}\right)^2\right.\\ &\left.+\frac{1}{2}\frac{I^+(2;x_r)}{p_r^+}\left(\frac{\partial'}{\partial x_r}\right) + \frac{3}{2}\left(\frac{I^+(1;x_r)}{p_r^+}\right)^2\left(\frac{\partial'}{\partial x_r}\right)\right)\ln(x_s-x_r+y_s)+\cdots\\ &\equiv \ln(x_r-x_s-y_s) + \sum_{m=1}^{\infty}(E^{\zeta_r})^m(D_r^{(m)})\ln(x_s-x_r+y_s). \end{aligned} \quad (A4)$$

Therefore, we find

$$\tilde{N}^{(r,s)}_{m,n} \equiv \tilde{N}^{(s,r)}_{n,m} = (D_r^{(m)})\tilde{N}^{(s,r)}_{n,0}. \quad (A5)$$

With the help of (A5), we can easily calculate $\tilde{N}^{(r,s)}_{m,n}$:

$$\tilde{N}^{(r,s)}_{1,1} = \frac{1}{(x_r-x_s)^2}, \quad (A6)$$

$$\tilde{N}^{(r,s)}_{2,1} = \frac{1}{(x_s-x_r)^3} + \frac{1}{(x_s-x_r)^2}\frac{I^+(1;x_r)}{p_r^+}, \quad (A7)$$

$$\tilde{N}^{(r,s)}_{2,2} = -\frac{3}{2}\frac{1}{(x_r - x_s)^4} + \frac{1}{(x_r - x_s)^2}\frac{I^+(1;x_r)}{p_r^+}\frac{I^+(1;x_s)}{p_s^+} + \frac{1}{(x_r - x_s)^3}\left(\frac{I^+(1;x_r)}{p_r^+} - \frac{I^+(1;x_s)}{p_s^+}\right), \tag{A8}$$

and

$$\tilde{N}^{(r,s)}_{3,1} = \frac{1}{(x_s - x_r)^4} + \frac{2}{(x_s - x_r)^3}\frac{I^+(1;x_r)}{p_r^+} + \frac{1}{2}\frac{1}{(x_s - x_r)^2}\frac{I^+(1;x_r)}{p_r^+} + \frac{3}{2}\frac{1}{(x_s - x_r)^2}\left(\frac{I^+(1;x_r)}{p_r^+}\right)^2, \tag{A9}$$

etc. Although (A6)–(A9) can be obtained from (2.11) after extremely tedious calculations, the reason why (A5) is valid for obtained results is not clear in such calculations.

[1] For reviews of string theory, see M. B. Green, J. H. Schwarz, and E. Witten, *Superstring Theory Vol. 2* (Cambridge U. P., New York, 1987).
[2] S. Mandelstam, Nucl. Phys. B **64**, 205 (1973).
[3] N. Berkovits, Nucl. Phys. B **304**, 537 (1988).
[4] J. L. Torres-Hernández, Phys. Rev. D **11**, 3565 (1975).
[5] For reviews on conformal field theory techniques in string theory, see, for example, D. Friedan, E. Martinec, and S. Shenker, Nucl. Phys. B **271**, 93 (1986); M. E. Peskin, "Introduction to string and superstring theory II," SLAC-PUB-4251 (1987).
[6] M. Ademollo, E. D. Giudice, P. Di Vecchia, and S. Fubini, Nuovo Cimento A **19**, 181 (1974). For vertex operator in the superstring, see K. Hornfeck, Nucl. Phys. B **293**, 189 (1987).
[7] D. J. Gross, J. A. Harvey, E. Martinec, and R. Rohm, Phys. Rev. Lett. **54**, 502 (1985).

Completeness relations for Maass Laplacians and heat kernels on the super Poincaré upper half-plane

Kazuto Oshima
Gunma College of Technology, Maebashi, Gunma 371, Japan

(Received 29 June 1989; accepted for publication 1 August 1990)

Simple completeness relations are proposed for Maass Laplacians. With the help of these completeness relations, correct heat kernels of (super) Maass Laplacians are derived on the (super) Poincaré upper half-plane.

I. INTRODUCTION

In the Polyakov string theory,[1,2] it is an important subject to compute determinants of Laplace-like operators (Maass Laplacians) on Riemann surfaces. Determinants are obtained from heat kernels by the ζ-function regularization method.[3]

The expressions for heat kernels of Maass Laplacians are given by Fay[4] on the Poincaré upper half-plane H. Heat kernels on Riemann surfaces are constructed by the Poincaré sum. In the recent paper by D'Hoker and Phong,[5] it is reported that Fay has pointed out to them that the discrete series that occurs in his expression for the heat kernel is erroneous and should be deleted. This remark is rather misleading. One may think that the discrete spectrum of the Maass Laplacian[4,6] does not contribute to the heat kernel. This is not true. The fact is that the error exists in the previous estimate with respect to the continuous spectrum. We discuss a procedure to obtain the correct heat kernel of the Maass Laplacian on H. A very simple and useful completeness relation is proposed.

It is also important in string theories to study Laplace-like operators on super Riemann surfaces. Aoki[7] obtained heat kernels of super Maass Laplacians on the super Poincaré upper half-plane $sH = \{(z,\theta)|\operatorname{Im} z > 0\}$. He solved heat kernel equations by a subtle trick. He used plane waves as eigenfunctions of the Maass Laplacian. The discrete spectrum was neglected, and there remains obscurity about the completeness of the eigenfunctions and about the short-time behavior of the heat kernel. The main purpose of the present paper is to make Aoki's argument more precise on the basis of the completeness relation mentioned above.

In Sec. II we discuss the procedure to derive the heat kernel of the Maass Laplacian in the bosonic case. In Sec. III we investigate the supersymmetric case. Section IV is devoted to a summary.

II. COMPLETENESS RELATIONS AND HEAT KERNELS ON H

Our starting point is the following spectral decomposition theorem:[4]

$$g(r) = \sum_{\{m\}} \frac{2|k|-2m-1}{4\pi} h(|k|-m) P_{|k|-m,k}(\operatorname{ch} r) + \frac{1}{8\pi i}\int_{\operatorname{Re} s=1/2} ds\, h(s) \frac{\sin 2\pi s}{\sin \pi(s+k)\sin \pi(s-k)} \times (s-\tfrac{1}{2}) P_{s,k}(\operatorname{ch} r), \tag{2.1}$$

or the completeness relation[8]

$$\frac{1}{2\pi}\delta(\operatorname{ch} r - 1) = \sum_{\{m\}} \frac{2|k|-2m-1}{4\pi} P_{|k|-m,k}(\operatorname{ch} r) + \frac{1}{8\pi i}\int_{\operatorname{Re} s=1/2} ds \frac{\sin 2\pi s}{\sin \pi(s+k)\sin \pi(s-k)} \times (s-\tfrac{1}{2}) P_{s,k}(\operatorname{ch} r), \tag{2.2}$$

where

$$\{m\} = \{m|m\in Z, 0\leqslant m < |k| - \tfrac{1}{2}\}, \quad k\in Z \text{ or } k\in Z+\tfrac{1}{2},$$

$$P_{s,k}(u) = \left(\frac{2}{1+u}\right)^s F\left(s-k, s+k, 1; \frac{u-1}{u+1}\right),$$

with F the hypergeometric function and $r = r(z_1, z_2)$ is the hyperbolic distance between z_1 and z_2 on H. Here,

$$\left(\frac{z-\bar{z}_0}{z_0-\bar{z}}\right)^{-k} P_{s,k}(\operatorname{ch} r(z,z_0))$$

is the eigenfunction of the Maass Laplacian

$$D_k = y^2(\partial_x^2 + \partial_y^2) - 2iky\partial_x, \tag{2.3}$$

with the eigenvalue $s(s-1)$ ($\operatorname{Re} s = \frac{1}{2}$, $s = |k| - m$).

The heat kernel $g^t{}_k(r)$ for the Maass Laplacian D_k is defined by setting $h(s) = e^{ts(s-1)}$ in (2.1). In Ref. 4 $g^t{}_k(r)$ is calculated to be

$$g^t{}_k(r) = \sum_{\{m\}} \frac{2|k|-2m-1}{4\pi} e^{t(|k|-m)(|k|-m-1)} \times P_{|k|-m,k}(\operatorname{ch} r) + \frac{\sqrt{2}e^{-t/4}}{(4\pi t)^{3/2}} \times \int_r^\infty db \frac{be^{-b^2/4t}}{\sqrt{\operatorname{ch} b - \operatorname{ch} r}} T_{2k}\left(\frac{\operatorname{ch}(b/2)}{\operatorname{ch}(r/2)}\right), \tag{2.4}$$

whereas the recent paper by D'Hoker and Phong[5] says Fay pointed out to them that (2.4) is not correct and the correct expression for $g^t{}_k(r)$ is given by

 0022-2488/90/123060-04$03.00

$$g^t{}_k(r)=\frac{\sqrt{2}e^{-t/4}}{(4\pi t)^{3/2}}\int_r^\infty db\,\frac{be^{-b^2/4t}}{\sqrt{\operatorname{ch} b-\operatorname{ch} r}}\,T_{2k}\left(\frac{\operatorname{ch}(b/2)}{\operatorname{ch}(r/2)}\right), \tag{2.5}$$

where T_{2k} is the $(2k)$th Chebyshev polynomial.

Equation (2.5) is different from (2.4) by the contribution of the discrete spectrum. This situation is likely to cause the misunderstanding that the discrete spectrum does not contribute to the heat kernel. The fact is that the sum of the contribution from the discrete spectrum and that from the continuous spectrum gives (2.5). In order to decide on this point, we derive (2.5) from (2.1).

Equation (2.4) was derived by misusing the formula [(22) in Ref. 4]

$$\frac{\sin 2\pi s}{\sin\pi(s-k)\sin\pi(s+k)}P_{s,k}(\operatorname{ch} r)$$
$$=-\frac{4}{\pi}\int_r^\infty db\,\frac{\operatorname{sh}(s-\frac{1}{2})b}{\sqrt{2\operatorname{ch} b-2\operatorname{ch} r}}\,T_{2k}\left(\frac{\operatorname{ch}(b/2)}{\operatorname{ch}(r/2)}\right),$$
$$|k|<\operatorname{Re} s<1-|k|, \tag{2.6}$$

in the second term of (2.1). For nonzero k, the line $\operatorname{Re} s=\frac{1}{2}$ lies out of the region (2.6) holds. To avoid this barrier we use the formula [(20) in Ref. 4]

$$\frac{\sin 2\pi s}{\sin\pi(s-k)\sin\pi(s+k)}P_{s,k}(u)$$
$$=-4(Q_{s,k}(u)-Q_{1-s,k}(u)), \tag{2.7}$$

where

$$Q_{s,k}(u)=-\frac{\Gamma(s-k)\Gamma(s+k)}{4\pi\Gamma(2s)}$$
$$\times\left(\frac{2}{1+u}\right)^s F\left(s-k,s+k,2s;\frac{2}{1+u}\right)$$
$$=-\frac{1}{4\pi}\left(\frac{2}{1+u}\right)^s$$
$$\times\sum_{n=0}^\infty\frac{\Gamma(s+k+n)\Gamma(s-k+n)}{\Gamma(2s+n)n!}\left(\frac{2}{1+u}\right)^n. \tag{2.8}$$

For $\operatorname{Re} s>|k|$, $Q_{s,k}(\operatorname{ch} r)$ has the following integral representation [(21) in Ref. 4]:

$$Q_{s,k}(\operatorname{ch} r)$$
$$=-\frac{1}{2\pi}\int_r^\infty\frac{db\,e^{(1/2-s)b}}{\sqrt{2\operatorname{ch} b-2\operatorname{ch} r}}T_{2k}\left(\frac{\operatorname{ch}(b/2)}{\operatorname{ch}(r/2)}\right),$$
$$\operatorname{Re} s>|k|. \tag{2.9}$$

We move the line of integration from $\operatorname{Re} s=1/2$ to $\operatorname{Re} s>|k|$ ($\operatorname{Re} s<1-|k|$) for $Q_{s,k}(Q_{1-s,k})$. Here, $(s-\frac{1}{2})Q_{s,k}$ has poles at $s=|k|-m$ ($m\in Z, 0\leqslant m<|k|-1/2$), and we find

$$-\frac{1}{2\pi i}\int_{\operatorname{Re} s=1/2}ds\,h(s)\left(s-\frac{1}{2}\right)Q_{s,k}$$
$$=-\frac{1}{2\pi i}\int_{\operatorname{Re} s>|k|}ds\,h(s)\left(s-\frac{1}{2}\right)Q_{s,k}+\sum_{\{m\}}h(|k|-m)\left(|k|-m-\frac{1}{2}\right)\operatorname{Res}(Q_{s,k})_{s=|k|-m}$$
$$=-\frac{1}{2\pi i}\int_{\operatorname{Re} s>|k|}ds\,h(s)\left(s-\frac{1}{2}\right)Q_{s,k}-\frac{1}{4\pi}\sum_{\{m\}}h(|k|-m)\left(|k|-m-\frac{1}{2}\right)P_{|k|-m,k}, \tag{2.10}$$

where

$$\operatorname{Res}(Q_{s,k})_{s=|k|-m}=-(1/4\pi)P_{|k|-m,k} \tag{2.11}$$

was used. Similarly we get

$$\frac{1}{2\pi i}\int_{\operatorname{Re} s=1/2}ds\,h(s)\left(s-\frac{1}{2}\right)Q_{1-s,k}$$
$$=\frac{1}{2\pi i}\int_{\operatorname{Re} s<1-|k|}ds\,h(s)\left(s-\frac{1}{2}\right)Q_{1-s,k}$$
$$-\frac{1}{4\pi}\sum_{\{m\}}h(|k|-m)\left(|k|-m-\frac{1}{2}\right)P_{|k|-m,k}. \tag{2.12}$$

Substituting (2.7), (2.11), and (2.12), into (2.1), we get

$$g(r)=\frac{1}{2\pi i}\left(\int_{\operatorname{Re} s>|k|}ds\,h(s)\left(s-\frac{1}{2}\right)Q_{s,k}(\operatorname{ch} r)\right.$$
$$\left.-\int_{\operatorname{Re} s<1-|k|}ds\,h(s)\left(s-\frac{1}{2}\right)Q_{1-s,k}(\operatorname{ch} r)\right). \tag{2.13}$$

The pole residues that occur in calculating the contribution of the continuous spectrum has canceled the contribution of the discrete spectrum. Paying attention to $h(s)=h(1-s)$ and using (2.9), we get the final formula:

$$g(r)=-\frac{1}{\pi i}\int_{\operatorname{Re} s>|k|}ds\,h(s)\left(s-\frac{1}{2}\right)Q_{s,k}(\operatorname{ch} r)$$
$$=\frac{\sqrt{2}}{4\pi^2 i}\int_{\operatorname{Re} s>|k|}ds\int_r^\infty db\,\frac{e^{(1/2-s)b}}{\sqrt{\operatorname{ch} b-\operatorname{ch} r}}h(s)$$
$$\times\left(s-\frac{1}{2}\right)T_{2k}\left(\frac{\operatorname{ch}(b/2)}{\operatorname{ch}(r/2)}\right). \tag{2.14}$$

Equation (2.2) can be read as

$$\frac{1}{2\pi}\delta(\operatorname{ch} r-1)$$
$$=-\frac{1}{\pi i}\int_{\operatorname{Re} s>|k|}ds\left(s-\frac{1}{2}\right)Q_{s,k}(\operatorname{ch} r)$$

$$= \frac{\sqrt{2}}{4\pi^2 i} \int_{\mathrm{Re}\, s > |k|} ds \int_r^\infty db \frac{e^{(1/2-s)b}}{\sqrt{\mathrm{ch}\, b - \mathrm{ch}\, r}} \times \left(s - \frac{1}{2}\right) T_{2k}\left(\frac{\mathrm{ch}\,(b/2)}{\mathrm{ch}\,(r/2)}\right); \quad (2.15)$$

$Q_{s,k}$ is regular for $\mathrm{Re}\, s > |k|$. If we set $h(s) = e^{ts(s-1)}$, the integration over s is a Gaussian. This Gaussian integral gives the correct expression of $g^t{}_k(r)$ (2.5) irrespective of the value $\mathrm{Re}\, s$. On account of this independence of the value $\mathrm{Re}\, s$, the same term appears in (2.4).

Equations (2.14) and (2.15) are very simple and useful, and they are used in the next section.

III. HEAT KERNELS ON *sH*

We study the heat kernel $\hat{G}^t{}_n(Z_1,Z_2)$ on sH defined by[7]

$$(\partial_t + \Delta_n)\hat{G}^t{}_n(Z_1,Z_2) = 0, \quad t>0, \quad (3.1)$$

$$\hat{G}^t{}_n(Z_1,Z_2) \overset{t\to 0}{\to} (z_{1\bar{1}}/-4)\delta(x_1-x_2)\delta(y_1-y_2) \times(\theta_1-\theta_2)(\bar{\theta}_1-\bar{\theta}_2), \quad (3.2)$$

where $Z=(z,\theta)\in sH$,

$$\Delta_n = z_{1\bar{1}}{}^2\partial_z\partial_{\bar{z}} + z_{1\bar{1}}\theta_{1\bar{1}}(\partial_z\hat{D}_- + \partial_{\bar{z}}\hat{D}_+) + (2n+1)z_{1\bar{1}}\hat{D}_-\hat{D}_+ - nz_{1\bar{1}}(\partial_z+\partial_{\bar{z}}) + 2n^2\theta_{1\bar{1}}(\hat{D}_- + \hat{D}_+) - n^2, \quad (3.3)$$

$$z_{ab} = z_a - z_b - \theta_a\theta_b, \quad z_{\bar{a}} = \overline{z_a}, \quad \theta_{ab} = \theta_a - \theta_b, \quad (3.4)$$

$$\hat{D}_+ = \partial_\theta + \theta\partial_z, \quad \hat{D}_- = \partial_{\bar{\theta}} + \bar{\theta}\partial_{\bar{z}}; \quad (3.5)$$

Δ_n is the Laplacian on superfields of weight n on sH. Then, $\hat{G}^t{}_n(Z_1,Z_2)$ can be expressed by invariant variables r, Δ, $\bar{\Delta}$ as

$$\hat{G}^t{}_n(Z_1,Z_2) = (z_{1\bar{2}}/z_{2\bar{1}})^n(g^t{}_n(r) + i\Delta\bar{\Delta}h^t{}_n(r)), \quad (3.6)$$

where

$$\mathrm{ch}\, r(Z_1,Z_2) = 1 - 2(z_{12}z_{\bar{1}\bar{2}}/z_{1\bar{1}}z_{2\bar{2}}), \quad (3.7)$$

$$\Delta = \frac{\theta_1 z_{2\bar{2}} + \theta_2 z_{\bar{2}1} + \theta_{\bar{2}} z_{12} + \theta_1\theta_2\theta_{\bar{2}}}{(z_{12}z_{2\bar{2}}z_{\bar{2}1})^{1/2}}, \quad (3.8)$$

and the complex conjugate one. The initial conditions (3.2) can be read as

$$g^t{}_n(r) \overset{t\to 0}{\to} 0, \quad (3.9)$$

$$\frac{h^t{}_n(r)}{\mathrm{sh}\, r} \overset{t\to 0}{\to} -\frac{1}{4\pi}\delta(\mathrm{ch}\, r - 1). \quad (3.10)$$

[This initial condition of $h^t{}_n(r)/\mathrm{sh}\, r$ is different from that of Ref. 7 by the factor of 1/2. It seems that the factor of 1/2 was missed in Ref. 7 in changes of variables.]

Aoki deduced $g^t{}_n(r)$ and $h^t{}_n(r)$ from $\hat{g}^t{}_n(z_1,z_2)$ and $\hat{F}^t{}_n(z_1,z_2)$ that are defined by

$$\hat{G}^t{}_n(Z_1,Z_2) = \hat{g}^t{}_n(z_1,z_2) + [\theta_1\bar{\theta}_1/(z_1-\bar{z}_1)]\hat{F}^t{}_n(z_1,z_2) + \text{terms involving } \theta_2,\bar{\theta}_2. \quad (3.11)$$

Comparing (3.6) and (3.11), we see

$$\text{the body of } g^t{}_n(r) = ((z_1-\bar{z}_2)/(z_2-\bar{z}_1))^{-n}\hat{g}^t{}_n(z_1,z_2), \quad (3.12)$$

$$\text{the body of } \frac{h^t{}_n(r)}{\mathrm{sh}\, r} = -\frac{1}{2}\left(\frac{z_1-\bar{z}_2}{z_2-\bar{z}_1}\right)^{-n}\hat{F}^t{}_n(z_1,z_2). \quad (3.13)$$

For $\hat{g}^t{}_n(z_1,z_2)$ and $\hat{F}^t{}_n(z_1,z_2)$, (3.1) and (3.2) reduce to

$$[\partial_t - (D_{-n} + n^2)]\hat{g}^t{}_n(z_1,z_2) + (2n+1)\hat{F}^t{}_n(z_1,z_2) = 0, \quad t>0, \quad (3.14)$$

$$(2n+1)D_{-n}\hat{g}^t{}_n(z_1,z_2) + [\partial_t - (D_{-n} + (n+1)^2)]\hat{F}^t{}_n(z_1,z_2) = 0, \quad t>0, \quad (3.15)$$

$$\hat{g}^t{}_n(z_1,z_2) \overset{t\to 0}{\to} 0, \quad (3.16)$$

$$\hat{F}^t{}_n(z_1,z_2) \overset{t\to 0}{\to} ((z_1-\bar{z}_2/(z_2-\bar{z}_1))^n y_1{}^2 \times\delta(x_1-x_2)\delta(y_1-y_2), \quad (3.17)$$

where D_{-n} is the Maass Laplacian in (2.3). Also, $\hat{g}^t{}_n(z_1,z_2)$ should obey

$$\{[\partial_t - (D_{-n} + (n+1)^2)][\partial_t - (D_{-n} + n^2)] - (2n+1)^2 D_{-n}\}\hat{g}^t{}_n(z_1,z_2) = 0, \quad t>0, \quad (3.18)$$

$$[\partial_t - (D_{-n} + n^2)]\hat{g}^t{}_n(z_1,z_2) \overset{t\to 0}{\to} -(2n+1)((z_1-\bar{z}_2)/(z_2-\bar{z}_1))^n y_1{}^2\delta(x_1-x_2)\delta(y_1-y_2), \quad (3.19)$$

and (3.16).

We use the Laplace transform in t. For

$$\tilde{g}^\lambda{}_n(z_1,z_2) = \int_0^\infty dt\, e^{-\lambda t}\hat{g}^t{}_n(z_1,z_2), \quad (3.20)$$

(3.16), (3.18), and (3.19) yield

$$\{[\lambda - (D_{-n} + (n+1)^2)][\lambda - (D_{-n} + n^2)] - (2n+1)^2 D_{-n}\}\tilde{g}^\lambda{}_n(z_1,z_2) = -(2n+1)((z_1-\bar{z}_2)/(z_2-\bar{z}_1))^n y_1{}^2 \times\delta(x_1-x_2)\delta(y_1-y_2). \quad (3.21)$$

To solve (3.21), Aoki used the plane-wave expansion for $\tilde{g}^\lambda{}_n(z_1,z_2)$ and the δ functions. Only the Whittacker functions were used as eigenfunctions of D_{-n}. The Whittacker functions correspond to the continuous spectrum. The La-

guerre polynomials that correspond to the discrete spectrum[6] were neglected. In this point, Aoki's argument is not perfect. We make Aoki's argument more precise. We expand $\tilde{g}^{\lambda}{}_{n}(z_1,z_2)$ and the δ functions by

$$((z_1-\bar{z}_2)/(z_2-\bar{z}_1))^n(s-\tfrac{1}{2})Q_{s,-n}(\mathrm{Re}\, s>|n|)$$

as in (2.14) and (2.15). After the inverse Laplace transform, we get

$$\hat{g}^{t}{}_{n}(z_1,z_2)=\frac{2n+1}{\pi i}\int_{c+iR}\frac{d\lambda}{2\pi i}\int_{\mathrm{Re}\, s>|n|}ds\frac{e^{t\lambda}(s-\frac{1}{2})((z_1-\bar{z}_2)/(z_2-\bar{z}_1))^nQ_{s,-n}(\mathrm{ch}\, u)}{(\lambda-(s+n)^2)(\lambda-(s-n-1)^2)}$$
$$=\frac{1}{2\pi i}\int_{\mathrm{Re}\, s>|n|}ds(e^{t(s+n)^2}-e^{t(s-n-1)^2})\left(\frac{z_1-\bar{z}_2}{z_2-\bar{z}_1}\right)^nQ_{s,-n}(\mathrm{ch}\, u), \tag{3.22}$$

where u is the body part of the super hyperbolic distance r in (3.7). Using the expression of $Q_{s,-n}$ (2.9), we find

$$\hat{g}^{t}{}_{n}(z_1,z_2)$$
$$=-\frac{((z_1-\bar{z}_2)/(z_2-\bar{z}_1))^n}{2\pi^{3/2}t^{1/2}}$$
$$\int_u^{\infty}db\frac{e^{-b^2/4t}\,\mathrm{sh}(n+\frac{1}{2})b}{\sqrt{2\,\mathrm{ch}\, b-2\,\mathrm{ch}\, u}}$$
$$\times T_{2k}\left(\frac{\mathrm{ch}\,(b/2)}{\mathrm{ch}\,(u/2)}\right). \tag{3.23}$$

Here, $\hat{g}^{t}{}_{n}(r)$ will be deduced as

$$\hat{g}^{t}{}_{n}(r)=-\frac{1}{2\pi^{3/2}t^{1/2}}\int_r^{\infty}db\frac{e^{-b^2/4t}\,\mathrm{sh}(n+\frac{1}{2})b}{\sqrt{2\,\mathrm{ch}\, b-2\,\mathrm{ch}\, r}}$$
$$\times T_{2k}\left(\frac{\mathrm{ch}\,(b/2)}{\mathrm{ch}\,(r/2)}\right). \tag{3.24}$$

The present result (3.24) just agrees with the result by Aoki (3.13) in Ref. 7. As we have seen in the bosonic case, it is possible to obtain the correct heat kernel by a rough treatment of the continuous spectrum and by neglect of the discrete spectrum. We see the same situation holds also for the present supersymmetric case. We have removed obscurity from the argument in Ref. 7 about obtaining $\hat{g}^{t}{}_{n}(r)$.

Then, $\hat{F}^{t}{}_{n}(z_1,z_2)$ is obtained from (3.14) and (3.24) as

$$\hat{F}^{t}{}_{n}(z_1,z_2)$$
$$=-\frac{1}{\pi i}\int_{\mathrm{Re}\, s>|n|}ds\left(\frac{(s-\frac{1}{2})(e^{t(s+n)^2}+e^{t(s-n-1)^2})}{2}\right.$$
$$\left.+\frac{(e^{t(s+n)^2}-e^{t(s-n-1)^2})}{4}\right)$$
$$\times\left(\frac{z_1-\bar{z}_2}{z_2-\bar{z}_1}\right)^nQ_{s,-n}(\mathrm{ch}\, u). \tag{3.25}$$

We see, from (2.15), $\hat{F}^{t}{}_{n}(z_1,z_2)$ satisfies condition (3.17) as t goes to 0. This fact, together with (3.13), will warrant the initial condition (3.10) for $h^{t}{}_{n}(r)$.

IV. SUMMARY

We have discussed the procedure to obtain the correct heat kernel on H. To calculate the contritibution from the continuous spectrum, it is inevitable to move the line of integration. The contribution from the discrete spectrum cancels the pole residues that originate in moving the line of integration. The very simple and useful completeness relation was given for the Maass Laplacian.

We have obtained the heat kernel on sH, with the help of the completeness relation mentioned above. We made Aoki's argument about solving the heat kernel equation on sH more precise. We also confirmed explicitly that the heat kernel satisfies the initial condition.

[1] A. M. Polyakov, Phys. Lett. B **103**, 207, 211 (1981).
[2] O. Alvarez, Nucl. Phys. B **216**, 125 (1983).
[3] E. D'Hoker and D. H. Phong, Nucl. Phys. B **269**, 205 (1986); Nucl. Phys. B **278**, 225 (1986); Commun. Math. Phys. **104**, 537 (1986).
[4] J. Fay, J. reine und angewende Math. **293**, 143 (1977).
[5] E. D'Hoker and D. H. Phong, Rev. Mod. Phys. **60**, 917 (1988).
[6] A. Comtet and P. G. Houston, J. Math. Phys. **26**, 185 (1985); A. Comtet, Ann. Phys. **173**, 185 (1987); C. Grosche, Ann. Phys. **187**, 110 (1988).
[7] K. Aoki, Commun. Math. Phys. **117**, 405 (1988).
[8] K. Oshima, Prog. Theor. Phys. **81**, 286 (1989).

Test fields on compact space-times

Ulvi Yurtsever[a)]
International Center for Theoretical Physics, Trieste, Italy

(Received 6 November 1989; accepted for publication 25 July 1990)

In this paper, some basic aspects of (Lorentzian) field theory on compact Lorentz manifolds are studied. All compact space-times are acausal, i.e., possess closed timelike curves; this makes them a useful testbed in analyzing some new notions of causality that will be introduced for more general acausal space-times. In addition, studying compact space-times in their own right raises a wide range of fascinating mathematical problems some of which will be explored. It will be shown that it is reasonable to expect Lorentzian field theory on a compact space-time to provide information on the topology of the underlying manifold; if this is true, then this information is likely to be "orthogonal" (or complementary) to the information obtained through the study of Euclidean field theory.

I. INTRODUCTION

As a fundamental physical theory, general relativity is well known for not imposing any constraints on the geometry of space-time other than the Lorentzian-manifold structure and the Einstein field equations. These constraints are mild: Every noncompact manifold admits a Lorentz metric, and unless one puts rather strong (energy) conditions on the form of the admissible stress-energy tensors, most any Lorentz metric is allowed as a possible solution. In specific problems, it may be appropriate to impose additional model-dependent constraints (such as symmetry or asymptotic conditions) on the geometry, and more generally, under a wide range of physical circumstances it is reasonable to assume the classical energy conditions. However, it is neither suggested nor warranted by the theory to discard any entire class of space-times as "unphysical," regardless of how strange and counterintuitive their properties may be. Attitudes that lead to such selective, *ad hoc* dismissals of space-time phenomena may be misleading and counterproductive; consider, for an example, the history of precisely these kind of attitudes that were held against singularities and horizons in the early decades of relativity.

Currently, causality conditions are widely believed to be natural constraints to impose on realistic space-times; e.g., a physically admissible space-time is generally assumed to be free of closed causal curves. The standard arguments given to justify this view all seem to be based on the notion of free will (see, e.g., Ref. 1, p. 189). The mathematical embodiment of this viewpoint, at least in classical general relativity, is the much more general strong cosmic censorship hypothesis (CCH) (Penrose,[2] Clarke *et al.*[3]). The "mildest" acausal space-times are those that simply fail to be globally hyperbolic, i.e., which violate only the strongest of all causality conditions. Such space-times need not possess any closed causal curves. The strong CCH asserts that space-times that develop from regular, well posed, generic initial data never violate global hyperbolicity, i.e., that the maximal Cauchy developments of such data are inextendible. This is equivalent to the assertion that all Cauchy horizons that develop from arbitrary (but well posed and regular) initial data are unstable (nongeneric). Specific examples of such horizons are obtained when the development of the Cauchy data gives rise to closed causal curves, lying in a region to the future of the initial surface. The null boundary that separates this region from the remaining, causal part of space-time is a Cauchy horizon, and the strong CCH predicts that this horizon will be unstable against small perturbations, thus preventing one from ever creating the closed causal curves that lie in its future.

Recently, it has been discovered[4] that it is, in fact, *possible* to prepare regular, generic Cauchy data whose unique (stable) evolution produces closed timelike curves, provided one is allowed to use (i) "exotic" matter fields that violate, among other features of classical, ordinary matter, the averaged weak energy condition,[5,4] and (ii) an initial Cauchy surface with wormhole topology [e.g., with topology (R^3 with a handle) $\cong R^3 \# (S^2 \times S^1)$ in the asymptotically flat case, or (S^3 with a handle) $\cong S^3 \# (S^2 \times S^1)$ in the compact case]. Indeed, with the example that is described in detail in Ref. 4, we explicitly construct a Cauchy horizon that almost certainly (the rigorous proof being not yet published) *is* stable, and which thus constitutes a counterexample to the strong CCH as described above. In this construction, exotic matter fields (that violate the averaged weak energy condition) play a crucial role, not only in maintaining a traversible wormhole over macroscopic timescales, but also in providing for the stability of the Cauchy horizon (see Ref. 4 for details).

The responsibility of ruling out closed causal curves in general relativity thus rests with quantum field theory (which *might* rule out the exotic stress-energy tensors of the kind needed for the above construction), and quantum gravity (which *might* prohibit the topology change that is necessary to create Cauchy surfaces with wormhole topology, when the spatial sections of space-time are initially simply connected). Given the nature of these two fields of research, it is reasonable to assume that the answers to the problems above are not going to be firmly in hand any time soon. Moreover, the strong-CCH argument can only rule out those closed causal curves which are "produced" when the space-time initially does not have them; it cannot rule out

[a)] Present address: The Enrico Fermi Institute, University of Chicago, Chicago, Illinois 60637.

 0022-2488/90/123064-15$03.00

those causal loops that may be "primordial," i.e., which may exist in the space-time eternally: Even if the strong CCH were to be proven "true" eventually by definitive results in quantum gravity or field theory, this would still leave the question open for space-times with primordial closed causal curves.

These considerations have recently led the author and his collaborators (Friedman *et al.*[6]) to the conclusion that it is important to explore the alternative possibility, namely, that the laws of physics might *not* rule out closed causal curves. In the course of their investigations, the authors have arrived at a mathematically well-defined viewpoint that allows theoretical physics to continue consistently in the presence of causal loops; this is embodied in the Principle of Self Consistency which we describe briefly in Sec. III A below. We will not expound this viewpoint any further here as it and its consequences are discussed in detail in a separate publication.[6] Instead, in this paper we will study some basic aspects of (Lorentzian) field theory on compact Lorentz manifolds, which we hope will provide useful insights into the physics and mathematics of more general acausal space-times.

All compact Lorentz manifolds possess closed timelike curves (see Sec. II below), and compactness allows elegant mathematical formulations for many of the notions that we will introduce below for more general acausal space-times (Sec. III). In addition, we believe there are good mathematical reasons to study compact space-times in their own right. It is well known that an astonishing amount of information on the topology and geometry of a compact (Riemannian) manifold can be obtained by studying Euclidean field theory on its background. Examples are the analysis of the Laplacian and other related elliptic operators on Riemannian manifolds (reviewed in Ref. 7), the more recent advances in four-dimensional topology achieved through the study of (Euclidean) Yang–Mills fields,[8] and the recent application of Euclidean quantum-field-theory ideas to the analysis of the topology of (compact) low-dimensional manifolds.[9,10] We will see that it is reasonable to expect Lorentzian field theory on a compact space-time to provide similar kinds of information on the topology of the underlying manifold; if so, it is likely that such information will be "orthogonal" (or complementary) to the information obtained through the study of Euclidean field theory. Furthermore, there are too many serious problems with the standard "Wick-rotation" argument to justify Euclidean field theory in general, in an arbitrary curved space-time (see Ref. 11 and the references therein). It is our view that ultimately all path integrals in field theory must be computable (after suitable regularization) in the Lorentzian regime, and studying Lorentzian field theory on compact spacetimes may offer fresh clues as to how such computations could be done.

Our most general problem, then, is the study of (Lorentzian) Yang–Mills fields on compact space-times, possibly also coupled to (spinorial) fermions. In this generality, the problem raises issues whose resolution is likely to involve long and extensive investigations and a substantial research effort. As a prelude to such investigations, in this paper we will study simpler (linear) field theories involving differential forms as test fields on a compact background space-time (see Sec. V for details). These theories are simple mathematical generalizations of Maxwell's theory. In fact, for zero-forms in two dimensions our theory reduces to that of a scalar field ϕ satisfying $\Box\phi = 0$, and for one-forms in four dimensions it is the standard Maxwell theory [Abelian Yang–Mills for a trivial $U(1)$-bundle over space-time]. We will start, in Sec. II, with a brief review on compact space-times. In Sec. III we introduce some new notions of causality (for acausal space-times) which will motivate most of the discussion that follows. These notions are in turn motivated by the Principle of Self-Consistency which is described briefly in Sec. III A. In Sec. IV, we illustrate our causality notions by studying two-dimensional compact space-times where our analyses become particularly simple. Section V describes the mathematical details of the general formalism with differential forms as test fields. In Sec. VI, we state some geometric criteria related to the causality notions of Sec. III, these criteria involve the geometry of null hypersurfaces and the spectrum of the d'Alembertian on-forms. Section VII contains a brief discussion on the problems of quantizing a classical field theory defined on a compact background space-time.

The primary goal of this paper is to communicate to the mathematical–physics community (and to some extent also to topologists and geometers) the wide range of fascinating questions on compact space-times that still remain largely unexplored. Detailed formulation and/or proof of some of the results (notably in Sec. VI) will be omitted from the discussion that follows. These can be found, along with a number of conjectures on questions left unresolved in this paper, in Ref. 12.

II. COMPACT SPACE-TIMES

By a *space-time* we will always mean a real, connected, C^∞ manifold M without boundary, with arbitrary dimension $n(\geqslant 2)$, and with a smooth (C^∞) metric g of Lorentz signature $(+,+,\ldots,+,-)$. For any smooth manifold the existence of a Lorentz metric is equivalent to the existence of a global line field (a smooth one-dimensional distribution); see Markus.[13] Every noncompact manifold admits a global line field; in fact, every noncompact manifold admits a vector field without zeros. (There always exists a vector field with isolated zeros; by noncompactness, it is possible to push the set of zeros to "infinity" using a sequence of diffeomorphisms that locally approach identity around every point.) For a *compact*, orientable manifold M, the existence of a Lorentz metric and the following conditions are all equivalent: (i) There exists a global line field on M; (ii) $\chi(M) = 0$ [$\chi(M)$ denotes the Euler number of M]; (iii) M admits a nowhere-zero vector field (not necessarily timelike). To prove (i) $\Rightarrow$ (ii), consider the double covering $\tilde{M}$ of M associated with the given line field. By construction, there always exists a nowhere-zero vector field on $\tilde{M}$; this implies $\chi(\tilde{M}) = \chi(M) = 0$ (see, e.g. Hirsch,[14] Theorem 5.2.2, p. 133). The Euler number of a compact (orientable) manifold M is equal to the Euler characteristic of its tangent bundle, and it is a standard result that every vector bundle over M with vanishing Euler characteristic admits a global (nowhere-zero) section (Hirsch,[14] Theorem 5.2.10, p. 137);

this proves (ii) $\Rightarrow$ (iii). Finally, given a nowhere-zero vector field X on M, let $\hat{X} \equiv X/[g_R(X,X)]^{1/2}$, and let ω be the one-form equivalent to $\hat{X}$ with respect to some Riemannian metric g_R on M. Then $g_L \equiv g_R - 2\omega \otimes \omega$ is a smooth metric of Lorentz signature; thus M admits a Lorentz metric whenever (iii) holds. A space-time (M,g) is called *time orientable* if the bundle of timelike vectors over M is disconnected (with two components). Time orientability is equivalent to the existence of a nowhere-zero *timelike* vector field on M [thus (M,g) is time orientable when M is simply connected]. If (M,g) is not time orientable, then its double covering $(\tilde{M},\pi^* g)$ (the same covering as that associated with the timelike line field on M, $\pi:\tilde{M} \to M$ being the projection) is a connected, time-orientable space-time. Thus there is no real loss of generality in assuming that all space-times we consider are both oriented and time orientable, and we will make this assumption unless stated otherwise.

Every odd-dimensional compact (orientable) manifold has zero Euler number, but for even-dimensional compact manifolds, $\chi(M) = 0$ is a nontrivial topological obstruction to the existence of Lorentz structures. Since we focus mostly on even-dimensional compact space-times in this paper (for reasons that are described in Sec. V below), we will now discuss this obstruction in a bit more detail.

In two dimensions, $\chi(M) = 0$ completely fixes the topology of a compact space-time to be that of the two-torus (Sec. IV). For a compact four-manifold M, combining Poincaré duality[15] with $\chi(M) = 0$ gives $b_1 = 1 + b_2/2 \geqslant 1$, where $b_k(M)$ denote the Betti numbers of M. Hence, a compact, four-dimensional space-time (M,g) cannot be simply connected.[1] It follows that (M,g) has a noncompact, (universal) simply connected covering space $(\tilde{M},\tilde{g})$; this might lead one to the conclusion that $(\tilde{M},\tilde{g})$ is a more "natural" model for the space-time than (M,g) itself (Ref. 1, p. 190). We will not subscribe to this viewpoint, but instead adopt the view that space-time topology is fixed *a priori*, analogously to fixing the equations that physical fields satisfy. From this point of view, the statement that space-time has the topology of M is equivalent to imposing certain periodicity constraints on all physical fields (including the metric) defined on the covering space $\tilde{M}$.

In (even) dimensions higher than four, the obstruction $\chi(M) = 0$ becomes milder. In particular, there exist higher-even-dimensional simply connected compact space-times. A six-dimensional example is $S^3 \times S^3$, equipped with the Lorentz structure consisting of the canonical Lorentz metric on S^3 direct summed with the canonical Riemannian metric on S^3. [By the *canonical* Lorentz metric on an odd sphere S^{2n-1}, we mean the following: With the imbedding $S^{2n-1} = \{|z_1|^2 + \cdots + |z_n|^2 = 1\} \subset C^n \equiv R^{2n}$, consider the nowhere-zero vector field X on S^{2n-1} associated with the one-parameter flow $\phi_t:(z_1,\ldots,z_n) \mapsto (e^{it}z_1,\ldots,e^{it}z_n)$. In the standard coordinates on R^{2n}, X is given by the vector field

$$\sum_{k=1}^{n} \left(-y_k \frac{\partial}{\partial x_k} + x_k \frac{\partial}{\partial y_k} \right)$$

tangent to S^{2n-1}. Let ω be the one-form equivalent to the (unit) vector field X with respect to the canonical Riemannian metric g_R on S^{2n-1}. Then the canonical Lorentz structure on S^{2n-1} is defined by the metric $g_L \equiv g_R - 2\omega \otimes \omega$.] In fact, in all even dimensions including four, any manifold M with $\chi(M) \neq 0$ can be turned into a manifold with $\chi = 0$ after connected summing with certain fixed, well-known manifolds. For example, in four dimensions, consider the compact orientable manifolds $U_4 \equiv CP^2$ and $D_4 \equiv CP^2 \# T^4$. We have $\chi(U_4) = 3$ and $\chi(D_4) = 1$, where more generally $\chi(CP^m) = m + 1$, $\chi(T^n) = 0$, and we have used the well-known fact $\chi(M_1 \# M_2) = \chi(M_1) + \chi(M_2) - 2$ (Ref. 15). If M is a four-manifold with $\chi(M) = -p (p \geqslant 0)$, then $M \# (U_4 \# \cdots \# U_4)$ (where the connected sum in parenthesis has p terms) has zero Euler number; similarly, if $\chi(M) = +q (q \geqslant 0)$, then the Euler number of $M \# (D_4 \# \cdots \# D_4)$ (where the connected sum in parenthesis has q terms) vanishes. Therefore, $\chi = 0$ is not an essential restriction from the viewpoint of topological classification: Any information on the topology of four-manifolds M with $\chi(M) = 0$ directly translates into equally valuable information on the topology of all orientable compact four-manifolds. In higher $(2m)$ dimensions, it is possible to find similar well-known manifolds U_{2m} and D_{2m} with $\chi(U_{2m}) = 3$, $\chi(D_{2m}) = 1$, so that the above argument continues to apply. In particular, when $m = 2k$ one can choose $U_{4k} \equiv CP^{2k} \# (T^{4k} \# \cdots \# T^{4k})$ and $D_{4k} \equiv U_{4k} \# T^{4k}$, where the connected sum in parenthesis has $(k-1)$ terms.

When combined with extra assumptions on the topology of M, $\chi = 0$ may become a powerful restriction. As an example, consider the assumption that $\pi_1(M)$, the fundamental group of M, is Abelian.[16] It is known[17] that for the Betti numbers of a compact manifold with Abelian $\pi_1(M)$ the inequality $b_1(b_1 - 1) \leqslant 2b_2$ holds. If furthermore the constraint $\chi(M) = 0$ is imposed (which implies $b_1 = 1 + b_2/2$), then a four-manifold can only have the b_2 values 0, 2, 4, and 6. Therefore, a compact four-dimensional space-time with Abelian fundamental group can only have the homology types (1,0), (2,2), (3,4), or (4,6) for its Betti numbers (b_1,b_2) (see Ref. 16 for details). For a six-dimensional compact space-time with Abelian $\pi_1(M)$, the same argument yields the weaker estimate $2b_1 \leqslant 1 + \sqrt{1 + 8b_2}$ on the first Betti number b_1.

All compact space-times are acausal; i.e., they contain closed causal (in fact timelike) curves. [By a "curve" we will always mean a C^1 curve. Note that our use of the term "acausal" for space-times with closed causal curves should not be confused with the notion of an acausal (or achronal) *subset* in space-time: A subset S is called acausal (achronal) if no two points of S can be joined by a causal (timelike) curve.] This is stated (but not quite proved) as Proposition 6.4.2 in Ref. 1. For a proof, consider the covering of a compact space-time (M,g) by open sets of the form $\{I^+(p), p \in M\}$. Since M is compact, this covering has a finite subcover $\{I^+(p_1),\ldots,I^+(p_k)\}$. The point p_1 is contained in $I^+(p_{l_1})$ for some $1 \leqslant l_1 \leqslant k$; the point p_{l_1} is contained in $I^+(p_{l_2})$, and so on. Since the subcover is finite, eventually some point p_{l_r} must belong to $I^+(p_{l_s})$ where $s \leqslant r$. Then there is a past-directed timelike curve joining p_{l_s} to p_{l_r} (since $s \leqslant r$), and another one joining p_{l_r} to p_{l_s} [since $p_{l_r} \in I^+(p_{l_s})$];

this gives a closed timelike curve through p_{I_+} (and p_{I_-}) in M. In general, the (open) subset of points of (M,g) through which there are closed timelike curves (the "chronology-violating set" of M) can be written as the disjoint union of subsets of the form $I^+(p)\cap I^-(p)$ (Ref. 1, Proposition 6.4.1). For a compact space-time (M,g) the chronology violating set may or may not be equal to M; if there are closed timelike curves through every point and if in addition M is simply connected, then (M,g) does not admit any global spacelike hypersurfaces. [A *global spacelike hypersurface* is a closed, imbedded, $(n-1)$-dimensional spacelike submanifold without edge.] This is because such a hypersurface divides a simply connected (M,g) into two disconnected pieces (say M_1 and M_2), and a future-directed timelike curve can intersect the hypersurface only in the direction from M_1 to M_2 and not from M_2 to M_1. For example, $S^3\times S^3$ does not admit a global spacelike hypersurface. Even when M is *not* simply connected, a compact space-time may not admit such hypersurfaces; an example in four dimensions is $S^3\times S^1$ with the canonical Lorentz metric on S^3 and the Riemannian metric on S^1.

Another set of constraints under which a compact space-time does not admit global spacelike hypersurfaces is the strong energy condition coupled with the usual genericity conditions on the curvature tensor.[1] For the proof see Newman.[18] We will not impose these or any other restrictions on the curvature tensor of the compact Lorentz manifolds we consider: The energy conditions are irrelevant physically in the context of acausal space-times (see Sec. I), and because compact space-times do not admit an initial-value interpretation, genericity conditions on curvature are not as plausible as in the case of the space-times developing from well-posed Cauchy data. For similar reasons, we will not impose any extra topological constraints on M either; e.g., we will not demand that a spin structure exists on (M,g). It is well known[19] that a *noncompact* space-time admits a spin structure if and only if it is parallelizable. This result appears to be false in the compact case, and the necessary and sufficient conditions for a compact space-time to admit spin structures are not well understood (see Ref. 20 for a detailed discussion and references).

One significant geometric assumption that we will impose on our compact space-times is that of geodesical completeness. In contrast with the Riemannian case, a compact Lorentz space is not necessarily complete. Let $\gamma\colon [0,L)\to M$ be (a terminal segment of) an incomplete (timelike or spacelike) geodesic in (M,g), which is parametrized by the arclength: $g(\gamma_*,\gamma_*)=\pm 1$, and which is inextendible beyond the parameter value L. Since γ is a geodesic, it is "locally flat" topologically, i.e., in every normal neighborhood of any point p on it the curve γ looks like a straight line (a Minkowski-space geodesic). In particular, if g_R is any Riemannian metric on M, and U and V are normal neighborhoods around $p\in\gamma$ with respect to the metrics g_R and g, respectively, then in both (respective) normal coordinate systems on $U\cap V$ γ appears (nearly) straight. Since the curve $\gamma\colon [0,L)\to M$ is inextendible and M (being compact) is complete (as a metric space) in any Riemannian metric, this implies that γ has infinite length (although it is not necessarily a geodesic) with respect to the metric g_R. Therefore, $\gamma(s)$ (where s is the affine parameter) is essentially an "infinite" curve; it has finite length in the Lorentz metric g only because $\gamma_*(s)$ asymptotes an (infinitely large) null vector as $s\to L$ (Fig. 1). [This behavior is similar to the incompleteness behaviors of the Taub–NUT space-time[21] and its two-dimensional analog, the "Misner space".[22]] To express this claim more precisely, pick a unit timelike vector field X on (M,g) [$g(X,X)=-1$], and let $\omega\equiv g(X,\cdot)$. We claim that $|g(X,\gamma_*(s))|$ is unboundedly large as $s\to L$. For the proof, assume that

$$|g(X,\gamma_*)|\leqslant C, \tag{1}$$

for some constant $C>0$. By the construction of the one-form ω, $g_R\equiv g+2\omega\otimes\omega$ is a Riemannian metric on M. But Eq. (1) implies

$$g_R(\gamma_*,\gamma_*)\leqslant \pm 1+2C^2, \tag{2}$$

and Eq. (2) is in contradiction with the fact that $\gamma\colon [0,L)\to M$ has infinite length in the Riemannian metric g_R. Exactly the same argument (with ± 1 replaced by 0) shows that an incomplete null geodesic must similarly have infinite Riemannian length and a blowing-up tangent vector. It follows from these results that g can be turned into a complete Lorentz metric just by slightly "opening-up" its light cones. More precisely, if $f\in C^\infty(M)$ with $|f|\ll 1$ is any suitable function supported in a neighborhood of γ, then the curve γ has infinite length (and hence cannot be incomplete) with respect to the Lorentz metric $g'\equiv g\pm f^2\omega\otimes\omega$ (Fig. 1). Consequently, we have proved that for a compact (orientable) manifold M the set $\mathrm{Lor}_c(M)$ of all *complete* Lorentz metrics is dense as a subset in $\mathrm{Lor}(M)$, where $\mathrm{Lor}(M)$ denotes the space of all (time-orientable) Lorentz metrics on M topologized with the (L^2-) norm obtained from any Riemannian structure. It is therefore reasonable (at least during the course of this preliminary investigation) to focus attention on complete compact space-times, and we will do so throughout most of the paper.

Whether a compact space-time is geodesically complete under the assumption of flatness remains a challenging open

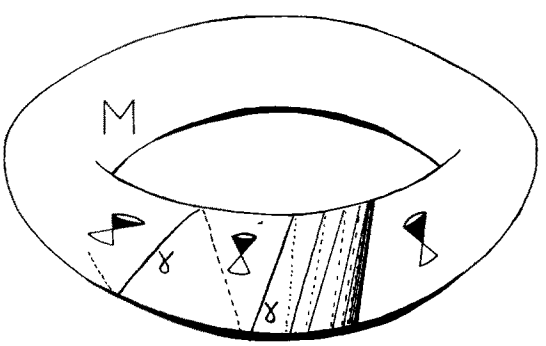

FIG. 1. Typical behavior of an incomplete timelike geodesic γ in a compact space-time M. The curve γ is infinitely long with respect to any Riemannian metric on M, but it has finite Lorentz length since the tangent vector γ_* asymptotes a null direction. In particular, by slightly opening up the null cones of M γ can be turned into a complete curve with infinite Lorentz length.

question. We conjecture that all flat compact space-times are complete; efforts to prove this conjecture are currently being investigated by the author. [A simple but intriguing example that is relevant to the issue of completeness is the two-dimensional Misner space.[22] For a brief description of Misner space see also p. 1448 of Ref. 4. The Misner space-time is flat, noncompact (with topology $S^1 \times R^1$), and geodesically incomplete; but all of its incomplete geodesics are totally future (or past) imprisoned in a fixed, compact subset.] Even with the assumption of completeness (which implies that the universal covering space is Minkowski), the classification of compact flat Lorentz spaces is a much more difficult problem than in the Riemannian case as it involves finding all discrete subgroups of the (noncompact) Lorentz group. This problem has only recently been solved (up to finite coverings) in Fried[23] and Goldman and Kamishima;[24] we refer the reader to these sources for further information.

III. CAUSALITY CONDITIONS FOR SPACE-TIMES WITH CLOSED CAUSAL CURVES

A. The principle of self-consistency

The only type of causality violation that is unacceptable in our view is the one embodied in the science-fiction concept of travelling back in time and killing one's grandmother before one's mother was conceived ("changing the past"). We shall enforce this view in the form of a *Principle of Self-Consistency*, which states that *the only solutions to the equations of physics that can occur locally are those which are globally self-consistent.* More precisely, the Principle of Self-Consistency (PSC) allows one to build a local solution to the field equations *only if* that local solution can be extended to be part of a (not necessarily unique) global solution that is well defined throughout the space-time.

Here, we adopt the viewpoint that in any space-time, including those that are globally hyperbolic, all solutions to the test-field equations of physics must ultimately (i.e., in principle) be globally well defined and smooth. Singularities that occur at points which belong to the space-time manifold are never realistic; in principle they must be removable either by refining the model (as in the treatment of a point charge in electromagnetism) or by enlarging the equations that define the test field (as in the treatment of shocks in hydrodynamics). [The only realistic singularities are those of general relativity, which may occur when the test-field equations are coupled with the Einstein equations. However, the occurence of GR singularities automatically implies the exclusion (or excision) of any singular "points" from the global space-time manifold, as the Lorentzian manifold structure of space-time must always be preserved. Throughout any singular space-time the metric and all other physical fields always remain smooth and globally self-consistent.]

It is only within the above viewpoint that we consider the PSC to be valid. Note also that the general formulation of the PSC conveniently leaves unspecified all relevant notions of smoothness such as "well defined," and "globally self-consistent." These notions will have to be made precise in accordance with the specific field theory that one chooses to study on the space-time. For our purposes in this paper, where we study simple linear theories for tensor fields, the standard notion of C^∞ smoothness is sufficient, although occasionally we will find it convenient to consider tensor fields that are smooth only in the (L^2-)distribution-sense (Sec. V). It is conceivable that distribution solutions will play a much more important role in this context when nonlinear field theories are treated.

In a globally hyperbolic space-time, the PSC is automatically enforced by the well-posed nature of the Cauchy problem (at least for those physical fields that satisfy hyperbolic equations). When space-time possesses closed causal curves, as a physical law the PSC by fiat forbids "changing the past," but in so doing it puts constraints on the allowed local solutions of the field equations; these, in turn, constrain the initial data for the Cauchy problem if the space-time admits a global spacelike hypersurface. Our main motive in the remainder of this paper will be the analysis of such constraints in the context of compact space-times.

B. Field-theoretical causality conditions

Let M be an n-dimensional compact manifold with zero Euler number. We will denote the vector space of smooth q-forms on M by $\Lambda^q(M)$. We will equip M with a fixed Riemannian structure g_R; then $\Lambda^q(M)$ naturally becomes a Hilbert space upon completion with respect to the inner product

$$\langle A,B \rangle_R \equiv \int_M A \wedge *_R B,$$

where $1 \leqslant q \leqslant n-2$, $A, B \in \Lambda^q(M)$, and $*_R$ denotes the Hodge dual with respect to g_R. If g is any Lorentz metric on M, then we will define the Hilbert-space structure on $\Lambda^0(M) = C^\infty(M)$ using the inner product

$$\langle f,g \rangle_R \equiv \int_M f \wedge *g = \int_M fg*1,$$

where $f, g \in \Lambda^0(M)$, and $*$ is the Hodge dual with respect to g. In Sec. V, we describe in detail the construction of field theories on the space-time (M,g) that involve as fields the elements of $\Lambda^q(M)$, $0 \leqslant q \leqslant n-2$. The field equations for these theories are simply the wave equation $\Box_Q A = 0$, $A \in \Lambda^q(M)$, where $\Box_Q$ is a second-order linear operator with hyperbolic symbol which is symmetric with respect to the Hilbert space structure $\langle , \rangle_R$ on $\Lambda^q(M)$. For $q = 0$, $\Box_Q$ reduces to the standard d'Alembertian operator on functions. When the dimension n of the space-time is even, the field theory of $(n/2-1)$-forms is especially interesting since it is invariant under conformal rescalings of the metric g (see Sec. V for details).

Let M be compact and even dimensional. Since (M,g) contains closed timelike curves, the PSC must constrain the solutions of the field equations that can be built locally in a given open neighborhood U in M. Those neighborhoods for which there are no constraints are special; we will call such neighborhoods *causally regular* since the presence of closed timelike curves cannot be detected by local observations confined in them.

Definition: An open subset U of M is called causally regular, if $\overline{U}$ (the closure of U) has an open neighborhood U_1 in M such that for every $(n/2-1)$-form A in $\Lambda^{n/2-1}(U_1)$ satisfying $\Box_Q A = 0$ on U there exists a C^∞ ex-

tension $\tilde{A}$ from U to M that satisfies the wave equation; i.e., if there exists an $\tilde{A}\in\Lambda^{n/2-1}(M)$ satisfying $\Box_Q\tilde{A}=0$ on M and $\tilde{A}\,|_U = A\,|_U$ on U.

Since the space-time M is trivially a causally regular (cr) neighborhood in M, we must be slightly more careful in defining the notion of a causally regular point.

Definition: A point $p\in M$ is called causally regular if *every* neighborhood around p contains a cr neighborhood of p.

Using the local solvability of the Cauchy problem for the wave equation, it is easy to see that if V is a convex normal neighborhood in M which is causally regular, then every open subset contained in V is a cr neighborhood. In particular, any point $p\in V$ is a cr point. It follows that every cr point p has a convex normal cr neighborhood around it. [If p is causally regular, then any convex normal neighborhood V_1 around p contains a cr neighborhood U of p. Let V be a convex normal neighborhood of p contained in U. Since U is contained in a normal neighborhood (V_1), solutions on V can be extended from V to U by the local well posedness of the Cauchy problem, and they can be extended from U to M by the causal regularity of the neighborhood U. Therefore, V is a convex normal cr neighborhood around p.] It also follows that every cr point has an open neighborhood consisting of cr points; hence the subset of all cr points in M is open. We will denote this open subset of M by C; the subset C coincides with M in a globally hyperbolic space-time.

Definition: A space-time (M,g) is called *benign* if $C=M$; (M,g) is called causally regular if C is nonempty.

For an even-dimensional compact space-time the above notions are conformally invariant. If (M,g) is a compact space-time with arbitrary (not necessarily even) dimension n, then our definitions can be generalized in the following way: Define an open subset U of M to be causally q regular (cqr), for $0\leqslant q\leqslant n-2$, if $\overline{U}$ has a neighborhood U_1 in M such that for every $A\in\Lambda^q(U_1)$ satisfying $\Box_Q A=0$ on U there exists an element $\tilde{A}$ in $\Lambda^q(M)$ which satisfies $\Box_Q\tilde{A}=0$ on M and $\tilde{A}\,|_U = A\,|_U$ on U. The remaining definitions naturally extend to cqr points, the open subset C_q, causally q-regular space-times, and q-benign space-times.

Another natural generalization of our causality notions can be obtained by considering, instead of smooth solutions to the wave equation, (L^2-)distributions [in the appropriate Hilbert space $\Lambda^q(M)$] lying in the domain of the adjoint $\Box_Q{}^*$ of $\Box_Q$ (see Sec. V). In this context, a distribution α satisfies the wave equation if the distribution $\Box_Q{}^*\alpha$ [defined by $\langle\Box_Q{}^*\alpha,\phi\rangle_R\equiv\langle\alpha,\Box_Q\phi\rangle_R$ for any C^∞ test form $\phi\in\Lambda^q(M)$] is in L^2 and is identically zero. Two distributions α,β coincide in an open subset $U\subset M$ if $\langle\alpha,\phi\rangle_R=\langle\beta,\phi\rangle_R$ for every smooth $\phi\in\Lambda^q(M)$ whose support is contained in U.

One obtains interesting variants of the field-theoretical causality conditions by introducing a *sheaf* Φ over (M,g) defined as the sheaf of germs of C^∞ local solutions to the wave equation. (See Chap. 5 of Ref. 25 for basic notions of sheaf theory.) Then one can define a point $p\in M$ to be *causally s-regular* if passing through every point of $\pi^{-1}(p)$ there is a global section of Φ, where $\pi:\Phi\to M$ is the projection. The space-time (M,g) is *s-benign* if there is a global section through every point of Φ. Although these definitions are not mathematically equivalent to the previous, field-theoretical ones (e.g., benign implies s-benign but not conversely), physically they capture the same notions of causality. Whether the sheaf-theoretical viewpoint can bring any significant insights into Lorentzian field theory on compact space-times remains to be seen; we will not pursue this viewpoint any further in this paper.

IV. CASE STUDY OF LORENTZ METRICS ON THE TWO-TORUS

For compact, two-dimensional orientable manifolds the Euler number is a complete topological invariant; i.e., all two manifolds are classified up to diffeomorphism by $\chi(M)$. It follows that for any compact two-dimensional space-time (M,g) M is diffeomorphic to the two torus T^2. In fact, in two dimensions any metric is locally conformally flat; it is then easy to prove that *almost every (in the sense of Lebesgue measure) compact, two-dimensional, complete space-time is globally conformal to a flat Lorentzian two-torus.* In the following few paragraphs we will first discuss some basic properties of flat Lorentz tori (with arbitrary dimensions), and later we will explore the causal behavior of two-dimensional compact space-times by studying scalar ($q=0$) field theory on 2-D toroidal backgrounds.

By a flat Lorentz torus we mean a complete, flat Lorentz metric on T^n. The universal covering space of a flat Lorentz torus is (R^n,η), the n-dimensional Minkowski space-time. It follows that any such torus can be obtained as a quotient $(R^n,\eta)/N$, where N is an n-dimensional lattice in (R^n,η); i.e., a discrete subgroup (consisting of translations) of the group of motions,

$$N=\{g|g:x\to x+m_1a_1+\cdots+m_na_n,\quad m_i\in Z\},\tag{3}$$

where $\{a_1,\ldots,a_n\}\subset(R^n,\eta)$ is a basis. Each of the vectors a_k may be timelike, null, or spacelike independently of the others; thus there exist many qualitatively different flat Lorentz tori in any dimension n. We will denote $(R^n,\eta)/N$, where N is given by Eq. (3), by the symbol $T_{(a_1,\ldots,a_n)}$.

The isometry group of $T_{(a_1,\ldots,a_n)}$ consists of the translations

$$x\to x+b,\quad b\in T_{(a_1,\ldots,a_n)},\tag{4}$$

and boosts $B\in O(n-1,1;R)\equiv L(n;R)$, which must satisfy

$$BL_{\{a_1,\ldots,a_n\}}=L_{\{a_1,\ldots,a_n\}},\tag{5}$$

where $L_{\{a_1,\ldots,a_n\}}\subset R^n$ denotes the lattice generated by integral linear combinations of $a_1,\ldots,a_n$. For $n\geqslant 3$, the conformal group of $T_{(a_1,\ldots,a_n)}$ is equal to the isometry group given by Eqs. (4)–(5). A two-dimensional flat torus $T_{(a_1,a_2)}$ has nonisometric conformal transformations if it admits nonconstant solutions to the wave equation; otherwise the isometry group and the conformal group of $T_{(a_1,a_2)}$ coincide.

How many distinct flat tori are there? It is easy to see that two flat Lorentz tori $T_{(a_1,\ldots,a_n)}$ and $T_{(b_1,\ldots,b_n)}$ are globally isometric if and only if

$$L_{\{Ba_1,\ldots,Ba_n\}}=L_{\{b_1,\ldots,b_n\}},\tag{6}$$

for some boost $B\in L(n;R)$. The tori $T_{(a_1,\ldots,a_n)}$ and $T_{(b_1,\ldots,b_n)}$ are globally *conformal* to each other if and only if

$$L_{\{Ba_1,\ldots,Ba_n\}}=L_{\{\lambda b_1,\ldots,\lambda b_n\}},\tag{7}$$

for some $B \in L(n;R)$ and $\lambda > 0$. For $n \geqslant 3$, any conformal diffeomorphism between $T_{(a_1,\ldots,a_n)}$ and $T_{(b_1,\ldots,b_n)}$ is in fact a homothety. In two dimensions, there exist nonhomothetic conformal diffeomorphisms between globally conformal flat tori iff the tori admit nonconstant solutions to the wave equation. For a fixed compact manifold M with zero Euler number, we will define the *conformal (Lorentzian) moduli space* of M as the space $\mathrm{Mod}(M) \equiv \mathrm{Lor}_c(M)/\sim$, where the equivalence relation $(M,g) \sim (M,g')$ holds iff there exists a diffeomorphism $\psi: M \to M$ which is conformal with respect to the Lorentz metrics g and g'; that is, iff there exists a ψ satisfying $\psi^* g' = \Omega^2 g$ for some positive $\Omega \in C^\infty(M)$. We will discuss the space $\mathrm{Mod}(T^2)$ in a little more detail shortly; it will turn out to be plausible to conjecture that $\mathrm{Mod}(T^2)$ has the structure of a finite-dimensional manifold except for singularities that form a subset of measure zero. For higher-dimensional manifolds, $\mathrm{Mod}(M)$ is in general an infinite-dimensional space.

The spectrum of the wave operator $\Box_Q$ on q-forms is easy to compute on a general flat torus $T_{(a_1,\ldots,a_n)}$. On the manifold T^n, we will fix an arbitrary *flat* Riemannian metric g_R. We let $\{\omega_1,\ldots,\omega_s\}$ [where $s = C(n;q)$] be any orthonormal basis (with respect to g_R) in $\Lambda^q(T_pM)$ consisting of eigenvectors of the operator $\hat{Q}_p^{(q)}: \Lambda^q(T_pM) \to \Lambda^q(T_pM)$ [with eigenvalues $\sigma_q(j): \hat{Q}_p^{(q)}\omega_j = \sigma_q(j)\omega_j$] at some point p in $M = T^n$. [For the definition of the operators $\hat{Q}_p^{(q)}$, ${Q^{(q)}}_p$, and $Q^{(q)}$ see Sec. V. All three operators reduce to identity for $q = 0$.] We will denote the q-forms on T^n which equal ω_j at p and which are parallel (constant) with respect to g_R (or equivalently with respect to g) by the same symbols ω_j. Since the metrics g_R and g are both flat, the q-forms ω_j satisfy $Q^{(q)}\omega_j = c\sigma_q(j)\omega_j$ throughout space-time, where $c > 0$ is a constant. Let $\{d_1,\ldots,d_n\} \subset (R^n,\eta)$ be the basis dual to $\{a_1,\ldots,a_n\}: \eta(d_i,a_j) = \delta_{ij}$. Then the (orthogonal) set of eigen-q-forms of $\Box_Q$ on $T_{(a_1,\ldots,a_n)}$ can be written in the form

$$\{\phi^j_{\{m_k\}}(x) \equiv \omega_j \exp[2\pi i \eta(m_1 d_1 + \cdots + m_n d_n, x)], \quad 1 \leqslant j \leqslant s, m_k \in Z\}, \tag{8}$$

where the coordinates $x \equiv (x_1,\ldots,x_n)$ belong to the standard global chart on (R^n,η), and the q-forms $\phi^j_{\{m_k\}}(x)$ are obviously well defined on the quotient $T_{(a_1,\ldots,a_n)} \equiv (R^n,\eta)/N$ [see Eq. (3)]. The complex exponential is a convenient way of grouping together the two linearly independent real eigenforms $\Re[\phi^j_{\{m_k\}}(x)]$ and $\Im[\phi^j_{\{m_k\}}(x)]$. According to Eq. (8),

$$S = \left\{ -4\pi^2 c\sigma_q(l) \sum_{i,j=1}^{n} m_i m_j \eta(d_i,d_j), \quad m_k \in Z, 1 \leqslant l \leqslant s \right\}, \tag{9}$$

where S denotes the spectrum of $\Box_Q$ acting on $\Lambda^q[T_{(a_1,\ldots,a_n)}]$.

We now return to discussing the causal properties of (generic) two-dimensional compact space-times. Since almost every two-dimensional (complete) (M,g) is globally conformal to some $T_{(a_1,a_2)}$, it follows that (i) there are closed timelike curves through *every* point of M, and (ii) (M,g) admits global spacelike hypersurfaces through every point. In fact, in two dimensions the $(q = 0)$ field theory of zero-forms is conformally invariant. Consequently, in order to understand the field-theoretical causality conditions of Sec. III for *generic* two-dimensional compact space-times, it is sufficient to analyze these conditions on *flat* Lorentz two-tori. We will carry out this analysis in the following remaining paragraphs of this section.

Consider a real scalar field ϕ satisfying $\Box\phi = 0 (\Box \equiv \Box_Q$ for $q = 0)$ on a flat toroidal background $T_{(a_1,a_2)}$. Initially, we will restrict our attention to those flat two-tori for which a_1 is spacelike and a_2 is a timelike vector orthogonal to a_1. Every such torus is globally conformal to a $T_{(1,r)} \equiv T_{[(1,0),(0,r)]}$, where $r > 0$ is a real number. Any solution of $\Box\phi = 0$ on $T_{(1,r)}$ naturally lifts to a solution of the wave equation on the covering space (R^2,η); that lifting (which we will denote by the same symbol ϕ) is one-periodic in the x coordinate and r-periodic in the t coordinate. Through every point of (R^2,η) there is a spacelike hypersurface of the form $\{t = \text{const}\}$; these define global spacelike hypersurfaces through every point of the quotient $T_{(1,r)}$. Let us fix (without any loss of generality) one such hypersurface $\{t = 0\}$, and consider the initial data $\{\phi_0(x), \dot{\phi}_0(x)\}$ that a solution of $\Box\phi = 0$ on $T_{(1,r)}$ induces on $\{t = 0\}$. By construction, $\phi_0(x)$ and $\dot{\phi}_0(x)$ are both periodic in x with period 1. In addition, the solution $\phi(x,t)$ on (R^2,η) that develops from the data $\{\phi_0(x), \dot{\phi}_0(x)\}$,

$$\phi(x,t) = \frac{1}{2}[\phi_0(x-t) + \phi_0(x+t)] + \frac{1}{2}\int_{x-t}^{x+t} \dot{\phi}_0(s)ds, \tag{10}$$

must be r periodic in t to pass to the quotient $T_{(1,r)}$. It is then easy to see from Eq. (10) that the initial data $\{\phi_0, \dot{\phi}_0\}$ induced from any solution of $\Box\phi = 0$ in $T_{(1,r)}$ are constrained in the following way:

$$\begin{aligned} \phi_0(x+1) - \phi_0(x) &= \dot{\phi}_0(x+1) - \dot{\phi}_0(x) \\ &= \phi_0(x+r) - \phi_0(x) \\ &= \dot{\phi}_0(x+r) - \dot{\phi}_0(x) = 0, \end{aligned} \tag{11}$$

$$\int_{\{t=0\}} \dot{\phi}_0(s)ds = 0. \tag{12}$$

The first set of constraints [Eqs. (11)] express the doubly periodic structure of ϕ_0 and $\dot{\phi}_0$ (with periods 1 and r), and the second [Eq. (12)] is an integral constraint on the data $\dot{\phi}_0$.

Consider first a two-torus $T_{(1,r)}$ with $r = p$; p being an integer. In this case, the only constraint on the initial data for ϕ is Eq. (12), since the periodicity constraints [Eqs. (11)] are automatically satisfied on the torus $T_{(1,p)}$. Consequently, every torus $T_{(1,p)}$ is benign: For any point $m \in T_{(1,p)}$, all sufficiently small convex normal neighborhoods V around m are cr neighborhoods, because the initial data induced by a local solution in V on a $\{t = \text{const}\}$ surface through m can always be extended (in many different ways) so as to satisfy the integral constraint Eq. (12).

Consider now a $T_{(1,r)}$ with $r = p/q$; p/q being a rational number with $(p,q) = 1$. In this case, Eqs. (11) constrain ϕ_0 and $\dot{\phi}_0$ to be $1/q$-periodic functions of x. Clearly, the two-

torus $T_{(1,p/q)}$ is benign: All normal neighborhoods V sufficiently smaller than $1/q$ in the x-direction are cr neighborhoods, because any local data $\{\phi_0,\dot{\phi}_0\}$ in V can be extended in a $1/q$-periodic manner while still satisfying Eq. (12).

Suppose now that r is an *irrational* number. In this case, the only solutions ϕ_0, $\dot{\phi}_0$ of the double-periodicity constraints Eqs. (11) are constants, and Eq. (12) implies that $\dot{\phi}_0 \equiv 0$: there are no nonconstant solutions to the wave equation on $T_{(1,r)}$. Clearly, a two-torus $T_{(1,r)}$ with irrational r is not benign.

Similar results hold for the two-tori $T_{(a_1,a_2)}$ where a_1 is spacelike but a_2 is an arbitrary vector. Every such torus is globally conformal to a $T_{[(1,0),(s,\tau)]}$, where $\tau > 0$. It is not very difficult to analyze the behavior of a scalar field ϕ satisfying $\Box\phi = 0$ on $T_{[(1,0),(s,\tau)]}$ using Eq. (10) and the initial surface $\{t = 0\}$, in a manner similar to the above analysis. It turns out that $T_{[(1,0),(s,\tau)]}$ is benign if and only if both $\tau - s$ and $\tau + s$ are rational: $\tau - s = p_1/q_1$, $\tau + s = p_2/q_2$, $(p_i,q_i) = 1$. The constraints on the initial data in this case are that ϕ_0, $\dot{\phi}_0$ be $1/m$-periodic, and

$$\int_{\{t=0\}} \dot{\phi}_0(s)\,ds = 0,$$

where $m \equiv [q_1,q_2]$ is the smallest common multiple of q_1 and q_2. If only one of $\tau - s$ or $\tau + s$ is rational, then $T_{[(1,0),(s,\tau)]}$ is not benign, but it still admits an infinite-dimensional space of solutions to the wave equation. The initial data of any such solution are constrained to satisfy $\dot{\phi}_0 = \pm\phi_0'$, and ϕ_0 must be $1/q$-periodic, where either $\tau - s$ or $\tau + s = p/q$, $(p,q) = 1$. When both $\tau - s$ and $\tau + s$ are irrational, the only global solutions to the wave equation on $T_{[(1,0),(s,\tau)]}$ are constants, consequently, $T_{[(1,0),(s,\tau)]}$ is not benign.

For a completely general flat two-torus $T_{[(s_1,\tau_1),(s_2,\tau_2)]}$, the analysis of causal regularity can be carried out using the observation that any solution ϕ to $\Box\phi = 0$ can be written in the form $\phi = f(u) + g(v)$ on the covering space (R^2,η), where $u \equiv t - x$, $v \equiv t + x$. The periodicity constraints for ϕ to pass to the quotient $T_{[(s_1,\tau_1),(s_2,\tau_2)]}$ imply that $f(u)$ be doubly periodic with periods $\tau_1 - s_1$ and $\tau_2 - s_2$, and $g(v)$ be doubly periodic with periods $\tau_1 + s_1$ and $\tau_2 + s_2$. It follows that $T_{[(s_1,\tau_1),(s_2,\tau_2)]}$ is benign if and only if both $(\tau_2 - s_2)/(\tau_1 - s_1)$ [or $(\tau_1 - s_1)/(\tau_2 - s_2)$ if $\tau_1 = s_1$] and $(\tau_2 + s_2)/(\tau_1 + s_1)$ [or $(\tau_1 + s_1)/(\tau_2 + s_2)$ if $\tau_1 = -s_1$] are rational.

We are now ready to answer the following natural question: What fraction of all (complete) Lorentz metrics on T^2 [or equivalently of all two-dimensional (complete) compact space-times] are benign? Clearly, it is sufficient to study this question on the (Lorentzian) conformal moduli space $\mathrm{Mod}(T^2)$: if g and g' are globally conformal metrics on T^2 then (T^2,g) is benign if and only if (T^2,g') is. What, then, does $\mathrm{Mod}(T^2)$ look like as a topological space?

In the Euclidean case, it is well known from elementary Riemann-surface theory that the conformal moduli space of Riemannian metrics on the torus T^2 is canonically isomorphic to the complex plane, i.e., it has the structure of a two-dimensional smooth manifold. In contrast, the Lorentzian moduli space $\mathrm{Mod}(T^2)$ has never been studied before, and it remains an interesting challenge to rigorously understand its structure. Some insight into that structure can be obtained by investigating a "component" of $\mathrm{Mod}(T^2)$, namely, the open subset consisting of all elements whose flat representatives are $T_{(a_1,a_2)}$, where a_1 is spacelike and a_2 is timelike. Conformally, these elements are all of the form $T_{[(1,0),\tau]}$, $\tau \in C^+$, where C^+ is the "upper half cone:" $C^+ \equiv \{\tau \in C \mid \Im\tau > |\Re\tau|\}$. After computations that follow the same logic as in the Riemannian case,[26] and using Eq. (7), one finds that two points τ' and $\tau = a + ib$ in C^+ represent the same element of $\mathrm{Mod}(T^2)$ iff

$$\tau' = [(\alpha + \beta a)^2 - \beta^2 b^2]^{-1}[\delta\alpha + (\beta\delta + \gamma\alpha)a + \gamma\beta(a^2 - b^2) + ib\,], \tag{13}$$

where

$$\begin{bmatrix} \alpha & \beta \\ \delta & \gamma \end{bmatrix} \tag{14}$$

is an admissible element of $\mathrm{SL}(2;Z)$. An element of $\mathrm{SL}(2;Z)$ of the form (14) is admissible to act on $\tau = a + ib$ if it satisfies the constraints

$$\gamma,\alpha > 0, \quad \gamma b > |\gamma a + \delta|, \quad |\alpha + \beta a| > |\beta b|. \tag{15}$$

Clearly, the open subset of $\mathrm{Mod}(T^2)$ that we are considering is C^+ with points identified under transformations of the form (13)–(15). However, the structure of this subset is complicated by the fact that neither $\mathrm{SL}(2;Z)$ nor any of its nontrivial subgroups act on C^+; the constraints (15) do not single out a subgroup, and what is worse, they depend on the argument τ.

Whatever its ultimate, detailed structure is, it seems quite clear that $\mathrm{Mod}(T^2)$ is two dimensional at least in the topological (but possibly not in the differentiable) sense. In particular, it is possible to introduce a canonical Lebesgue measure on $\mathrm{Mod}(T^2)$. It then follows from the results of the preceding paragraphs that the subset of all benign Lorentz metrics is of measure zero in $\mathrm{Lor}_c(T^2)$, because the points (s,τ) with rational coordinates form a subset of zero Lebesgue measure in the real plane R^2. Equivalently, among all complete Lorentz metrics on T^2 those that are benign form a nongeneric subset. In Sec. 6.3 of Ref. 12, we conjecture that this result is peculiar to two dimensions, and argue that for higher-dimensional compact manifolds M the Lorentz metrics that are benign probably form a generic subset in $\mathrm{Lor}_c(M)$.

We will close this section by briefly discussing the spectrum of the d'Alembertian $\Box$ on functions in a two-dimensional compact space-time. For brevity, we will restrict our attention to flat two-tori of the form $T_{[(1,0),(0,r)]}$, $r > 0$. (The properties of the spectrum that we will discuss below are conformally invariant, so it is sufficient to consider only flat two-tori.) We are interested in the following questions on the spectrum: (a) Is the zero-eigenspace (the kernel of $\Box$) infinite dimensional? (b) Is $\lambda = 0$ an accumulation point for the eigenvalues? (c) Can the eigenvalues have finite accumulation points on the real axis? (d) Are there any infinite-dimensional ($\lambda \neq 0$) eigenspaces? For an elliptic operator (e.g., the Laplacian on T^2), the answer to all four questions is no (see Sec. V for a more general discussion). For the

d'Alembertian $\Box$ on $T_{[(1,0),(0,r)]}$, the following information about the spectrum is easily derived using Eq. (9): When r is rational, the answer to all three questions (b), (c), and (d) is no, whereas the answer to (a) is yes. If r is irrational but r^2 is rational, then the answer to (a), (b), and (c) is no, and the answer to (d) is yes. If r and r^2 are both irrational, and r can be approximated abnormally well by rational numbers (a nongeneric property for irrational numbers), then the answer to (b), (c), and (d) is yes, and the answer to (a) is no. When r and r^2 are both irrational and r cannot be approximated abnormally well by rational numbers, the answer to (a), (b) is no, and the answer to (c) and (d) is yes.

V. THE GENERAL MATHEMATICAL FORMALISM

Let M be a compact orientable manifold with vanishing Euler number. We will choose a Riemannian metric g_R on M which will be kept fixed throughout the analysis of the compact space-times (M,g), where g is any Lorentz metric on M. All essential features of our analysis will be seen to be independent of the choice of the Riemannian metric g_R.

For $1\leqslant q\leqslant n-2$, $n\equiv\dim M$, the vector space of q-forms $\Lambda^q(M)$ has a natural Hilbert-space structure given by the completion of the (positive definite) inner product

$$\langle A,B\rangle_R\equiv\int_M A\wedge *_R B,\quad A,B\in\Lambda^q(M). \tag{16}$$

For $q=0$, we will define the Hilbert-space structure on $\Lambda^0(M)=C^\infty(M)$ using the inner product

$$\langle f,g\rangle_R\equiv\int_M f\wedge *g=\int_M fg*1,\quad f,g\in\Lambda^0(M), \tag{17}$$

where $*$ is the Hodge dual with respect to the Lorentz metric of the space-time (M,g). We will denote by the symbol H^q the Hilbert space obtained after completing $\Lambda^q(M)$ under the inner product(s) (16) or (17) (for $0\leqslant q\leqslant n-2$). The elements of H^q can be naturally identified with L^2 distributions (in the sense of q-forms) on M. By definition, any element $\alpha\in H^q$ is the limit of a Cauchy sequence $\{\alpha_n\}$ (with respect to the norm obtained from $\langle,\rangle_R$) of smooth q-forms in $\Lambda^q(M)$. When regarded as a distribution, the element α acts on a smooth test form $\phi\in\Lambda^q(M)$ through the relation

$$\alpha[\phi]\equiv\lim_{n\to\infty}\langle\alpha_n,\phi\rangle_R\equiv\langle\alpha,\phi\rangle_R. \tag{18}$$

On the infinite-dimensional vector space $\Lambda^q(M)$, $0\leqslant q\leqslant n-2$, we will introduce two more symmetric bilinear quadratic forms that depend on the Lorentz metric g on M; the form $\langle,\rangle_E$ given by

$$\langle A,B\rangle_E\equiv\int_M dA\wedge *dB,\quad A,B\in\Lambda^q(M), \tag{19}$$

and the (nondegenerate) form $\langle,\rangle_H$ given by

$$\langle A,B\rangle_H\equiv\int_M A\wedge *B,\quad A,B\in\Lambda^q(M). \tag{20}$$

The quadratic form $\langle,\rangle_H$ coincides with the inner product $\langle,\rangle_R$ for $q=0$. In fact, for any q the form $\langle,\rangle_H$ admits a unique, well-defined, nondegenerate (but indefinite) extension to the Hilbert space H^q:

$$\langle\alpha,\beta\rangle_H\equiv\lim_{n\to\infty}\langle\alpha_n,\beta_n\rangle_H,\quad \alpha,\beta\in H^q, \tag{21}$$

where $\{\alpha_n\}$ and $\{\beta_n\}$ are Cauchy sequences of smooth q-forms in $\Lambda^q(M)$ converging (in H^q) to α and β, respectively. It is not difficult to show that the right-hand side of Eq. (21) is independent of the choice of the sequences $\{\alpha_n\}$ and $\{\beta_n\}$. In contrast, the quadratic form $\langle,\rangle_E$ on $\Lambda^q(M)$ does not admit an extension to H^q since its definition involves differentiating the arguments A, $B\in\Lambda^q(M)$.

We will define the d'Alembertian operator $\Box:\Lambda^q(M)\to\Lambda^q(M)$ on q-forms as $\Box\equiv-\delta d$:

$$\Box_\alpha\equiv(-1)^{1+q+q(n-q)}*d*d\alpha,\quad \alpha\in\Lambda^q(M). \tag{22}$$

The operator $\Box$ is symmetric (formally self-adjoint) with respect to the indefinite quadratic form $\langle,\rangle_H$ on $\Lambda^q(M)$:

$$\langle\Box\alpha,\beta\rangle_H=\langle\alpha,\Box\beta\rangle_H,\quad \alpha,\beta\in\Lambda^q(M). \tag{23}$$

Furthermore, $\Box$ satisfies the property

$$\begin{aligned}\langle\alpha,\beta\rangle_E&\equiv\langle d\alpha,d\beta\rangle_H\\&=-\langle\Box\alpha,\beta\rangle_H\\&=-\langle\alpha,\Box\beta\rangle_H,\quad \alpha,\beta\in\Lambda^q(M),\end{aligned} \tag{24}$$

where $\langle,\rangle_E$ is defined by Eq. (19). [For a concise review of our conventions for Lorentzian exterior calculus see Sec. V of Ref. 12.]

We will define our field theory of q-forms by the action

$$S[A]\equiv\langle A,A\rangle_E=\int_M dA\wedge *dA,\quad A\in\Lambda^q(M), \tag{25}$$

where $0\leqslant q\leqslant n-2$. The (Euler–Lagrange) field equations for the theory (25) are simply $\Box A=0$. In fact, the first variation of the action (25) is

$$\delta S[A](\delta B)=-\langle\delta B,\Box A\rangle_H, \tag{26}$$

and the second variation is

$$\delta^2 S[A](\delta B,\delta C)=-\langle\delta B,\Box(\delta C)\rangle_H=\langle\delta B,\delta C\rangle_E, \tag{27}$$

which is independent of the q-form A.

For $q=0$, the wave operator $\Box$ is symmetric (formally self-adjoint) with respect to the Hilbert-space structure $\langle,\rangle_R$ on $\Lambda^0(M)$ [Eqs. (17) and (23)]; we will denote this symmetric operator by $\Box_Q:\Lambda^0(M)\to\Lambda^0(M)$. For $q\geqslant 1$, however, $\Box$ is symmetric with respect to the indefinite $\langle,\rangle_H$ but not with respect to the positive-definite $\langle,\rangle_R$. To cure this, we will introduce an operator $Q_p:T_pM\to T_pM$ via the relation

$$g_R(Q_pX,Y)=g(X,Y),\quad X,Y\in T_pM, \tag{28}$$

where $p\in M$ is any point. The operator Q_p is nonsingular, and symmetric with respect to g_R:

$$g_R(Q_pX,Y)=g(X,Y)=g(Y,X)=g_R(X,Q_pY). \tag{29}$$

We will denote the canonical extension of Q_p to the exterior algebra $\Lambda^q(T_pM)$ $(q\geqslant 1)$ by $\hat{Q}_p^{(q)}:\Lambda^q(T_pM)\to\Lambda^q(T_pM)$. Since the manifold M is orientable, there exists a positive $F\in C^\infty(M)$ that relates the volume forms Ω_R and Ω:

$$*_R 1=F*1,\quad F(x)>0,\quad F\in C^\infty(M). \tag{30}$$

We will define another operator $Q^{(q)}{}_p:\Lambda^q(T_pM)\to\Lambda^q(T_pM)$ by

$$Q^{(q)}{}_p \equiv [1/F(p)]\hat{Q}^{(q)}_p, \tag{31}$$

and let $Q^{(q)}:H^q \to H^q$ denote the unique extension of the operator $Q^{(q)}:\Lambda^q(M) \to \Lambda^q(M)$ [defined pointwise by Eq. (31)] to the entire Hilbert space H^q. Clearly, $Q^{(q)}$ is a linear, bounded [with norm $\|Q^{(q)}\| \leqslant \max_{p\in M}\{|\sigma_q(p)/F(p)|\}$, where $\sigma_q(p)$ is the largest (in absolute value) eigenvalue of $\hat{Q}^{(q)}_p:\Lambda^q(T_pM) \to \Lambda^q(T_pM)$], nonsingular (invertible), and local operator on H^q. [By $Q^{(q)}:H^q \to H^q$ being a local operator we mean the property that for any distribution $\alpha \in H^q$ that vanishes on an open $U \subset M$, $Q^{(q)}\alpha$ on U also vanishes. An $\alpha \in H^q$ vanishes on U if $\langle \alpha,\phi\rangle_R = 0$ for any test form $\phi \in \Lambda^q(M)$ whose support is contained in U.] It follows from Eqs. (29) and (31) that $Q^{(q)}$ is self-adjoint:

$$\langle Q^{(q)}\alpha,\beta\rangle_R = \langle \alpha,Q^{(q)}\beta\rangle_R, \quad \alpha,\beta \in H^q, \tag{32}$$

and that it satisfies

$$\langle Q^{(q)}\alpha,\beta\rangle_R = \langle \alpha,\beta\rangle_H, \alpha,\beta \in H^q. \tag{33}$$

For $1 \leqslant q \leqslant n-2$, we will define the wave operator $\Box_Q:\Lambda^q(M) \to \Lambda^q(M)$ by

$$\Box_Q \equiv Q^{(q)} \circ \Box. \tag{34}$$

The operator $\Box_Q$ is symmetric (formally self-adjoint) with respect to the Hilbert-space structure $\langle,\rangle_R$ on $\Lambda^q(M)$:

$$\begin{aligned}\langle Q^{(q)}\Box A,B\rangle_R &= \langle \Box A,B\rangle_H = \langle A,\Box B\rangle_H \\ &= \langle \Box B,A\rangle_H = \langle Q^{(q)}\Box B,A\rangle_R \\ &= \langle A,Q^{(q)}\Box B\rangle_R, \quad A,B \in \Lambda^q(M).\end{aligned} \tag{35}$$

Furthermore, it follows from Eqs. (24), (33), and (34) that

$$\begin{aligned}\langle A,B\rangle_E &= -\langle \Box_Q A,B\rangle_R \\ &= -\langle A,\Box_Q B\rangle_R, \quad A,B \in \Lambda^q(M),\end{aligned} \tag{36}$$

and, using Eq. (26),

$$\delta S[A](\delta B) = -\langle \Box_Q A,\delta B\rangle_R. \tag{37}$$

Consequently, the field equations $\Box A = 0$ for the action Eq. (25) can be written in the fully equivalent form $\Box_Q A = 0$, $A \in \Lambda^q(M)$.

When M is even dimensional, the quadratic form $\langle,\rangle_E$ on $\Lambda^{n/2-1}(M)$ is invariant under conformal rescalings of the metric g: If $g' = \Omega^2 g$, $\Omega > 0$, then the Hodge star $*:\Lambda^{n/2}(M) \to \Lambda^{n/2}(M)$ on half forms is the same operator for g' as it is for g, consequently, $\langle A,B\rangle_{E'} = \langle A,B\rangle_E$ for A, $B \in \Lambda^{n/2-1}(M)$. If $\psi:M \to M$ is a diffeomorphism and $g' = \psi^* g$, then it is easy to show that for all q, $\langle,\rangle_{E'} = \psi^*[\langle,\rangle_E]$, where

$$\psi^*[\langle,\rangle_E](A,B) \equiv \langle \psi^* A,\psi^* B\rangle_E, \quad A,B \in \Lambda^q(M). \tag{38}$$

Therefore, for general conformally related Lorentz metrics g' and g on M that satisfy $\psi^* g' = \Omega^2 g$, the quadratic forms $\langle,\rangle_{E'}$ and $\langle,\rangle_E$ on $\Lambda^{n/2-1}(M)$ are related by

$$\psi^*[\langle,\rangle_{E'}] = \langle,\rangle_E [\text{on } \Lambda^{n/2-1}(M)]. \tag{39}$$

In fact, it is known that the above results are also true in the converse direction: If g and g' are two Lorentz metrics on M with the same Hodge star $*:\Lambda^{n/2}(M) \to \Lambda^{n/2}(M)$ on half forms, then there exists a positive $\Omega \in C^\infty(M)$ such that $g' = \Omega^2 g$ (for a proof, see Dray *et al.*[27]). It follows that when $\langle,\rangle_{E'} = \langle,\rangle_E$ on $\Lambda^{n/2-1}(M)$ one similarly has $g' = \Omega^2 g$. In fact, if there exists a diffeomorphism $\psi:M \to M$ such that $\psi^*[\langle,\rangle_{E'}] = \langle,\rangle_E$ on $\Lambda^{n/2-1}(M)$, then $\psi^* g' = \Omega^2 g$ for some positive $\Omega \in C^\infty(M)$. Therefore, the quadratic form $\langle,\rangle_E$ on $\Lambda^{n/2-1}(M)$ is a *complete conformal invariant* for the space-time (M,g). In other words, if we let $Q(M)$ denote the space of all quadratic forms $\langle,\rangle_E$ on $\Lambda^{n/2-1}(M)$ that come from complete Lorentz metrics g, and we let $\sim$ denote the equivalence relation on $Q(M)$ defined by Eq. (39), then the quotient space $Q(M)/\sim$ is naturally isomorphic with the conformal moduli space $\mathrm{Mod}(M)$.

The adjoint $\Box_Q{}^*$ of the symmetric operator $\Box_Q:\Lambda^q(M) \to \Lambda^q(M)$ is defined as the operator with domain $D(\Box_Q{}^*) = \{\alpha \in H^q | \Box_Q{}^*\alpha \in H^q\}$, where $\Box_Q{}^*\alpha$ is the distribution whose action on a test form $\phi \in \Lambda^q(M)$ is given by $\Box_Q{}^*\alpha[\phi] \equiv \langle \alpha,\Box_Q\phi\rangle_R$. It is possible to prove that $\Box_Q$ is essentially self-adjoint on H^q. We will not describe this proof; but it proceeds in essentially the same way as the proof of essential self-adjointness of the Laplacian [by showing that $\pm i$ are in the resolvent set (after complexifying the Hilbert space H^q)], given, for example, in Sect. 11.7 of Richtmyer.[28] The differential operator $\Box_Q$ on $\Lambda^q(M)$ has hyperbolic symbol, therefore none of the regularity results for elliptic operators apply to $\Box_Q$. In particular, the spectrum of $\Box_Q$ is not necessarily discrete, nor the eigenspaces are necessarily finite dimensional. [By the "spectrum" of $\Box_Q$ we will always mean the point spectrum, i.e., the set of eigenvalues counted with multiplicity. For an elliptic operator L, any bounded infinite sequence $\{u_k\}$ for which $\{Lu_k\}$ is also bounded contains a Cauchy subsequence (Ref. 25, Theorem 6.6). This result by itself implies that the spectrum of L is "clean." It then follows from standard "elliptic regularity" (Ref. 25, Theorem 6.30) that the eigenvalues of L are smooth functions (as elements in the Hilbert space).] Since $\Box_Q$ is symmetric (in fact essentially self-adjoint) with dense domain $\Lambda^q(M)$ in H^q, it follows (i) that its eigenvalues are real, (ii) that the eigenspaces with distinct eigenvalues are orthogonal, and (since H^q is separable) (iii) that the (point) spectrum S of $\Box_Q$ is a countable subset in R (e.g., S cannot contain an open interval). It also follows that the eigenspaces of $\Box_Q$ span the Hilbert space H^q. About the eigenvectors of $\Box_Q$, we will make the following assumption: Every eigenspace admits an orthonormal basis consisting of smooth elements in $\Lambda^q(M)$; in particular, all vectors in a finite-dimensional eigenspace are smooth q-forms. We suspect that this assumption is true quite generally, but we will not attempt to prove it in this paper.

Let E_{λ_i} denote the eigenspace corresponding to the eigenvalue $\lambda_i \in R$ of $\Box_Q$. We will denote a complete orthonormal sequence of smooth eigenforms by $\{u_{ie}\}$, where i is an index for the distinct eigenvalues λ_i, and e indexes an orthonormal basis for the eigenspace E_{λ_i}. (For notational convenience, we will omit the superscripts q unless they are needed for clarity.) It follows from Eq. (36) that the quadratic form $\langle,\rangle_E$ on $\Lambda^q(M)$ is in diagonal form in the basis $\{u_{ie}\}$. The null space of the form $\langle,\rangle_E$ will be denoted by $N \subset \Lambda^q(M)$; N is equal to the zero-eigenspace E_0. The symbol H will denote the orthogonal projection $H:H^q \to N$. For-

mally, H can be written in the form

$$H=\sum_{i=0,e} u_{0e}\langle u_{0e}|, \tag{40}$$

where $\langle u_{ie}|$ denotes the bounded linear functional

$$\langle u_{ie}|(\alpha)\equiv\langle u_{ie},\alpha\rangle_R \tag{41}$$

on H^q. We will denote by I the image of $\Lambda^q(M)$ under $\Box_Q$:

$$I\equiv\Box_Q[\Lambda^q(M)]. \tag{42}$$

One has the orthogonal direct-sum decompositions

$$\Lambda^q(M)=N\oplus_R I=N\oplus_E I, \tag{43}$$

and the relations

$$\overline{I}=N^{\perp R},\quad \Lambda^q(M)=N^{\perp E}, \tag{44}$$

where overbar denotes closure (in H^q), and $\perp_R$, $\perp_E$ denote the orthogonal complements with respect to $\langle,\rangle_R$ *in* H^q, and $\langle,\rangle_E$ *in* $\Lambda^q(M)$, respectively. We will define the Green's operator $G:\Lambda^q(M)\to I$ by $G:\phi\mapsto G(\phi)$, where $G(\phi)$ is the unique solution in I of the equation

$$\Box_Q[G(\phi)]=\phi-H(\phi),\quad G(\phi)\in I. \tag{45}$$

Formally, G can be written in the form

$$G=\sum_{i\neq 0,e}\frac{1}{\lambda_i}u_{ie}\langle u_{ie}|. \tag{46}$$

It is possible to express the notion of causal q-regularity (Sec. III B) for an open neighborhood $U\subset M$ purely in terms of the quadratic form $\langle,\rangle_E$. Let $Z_U\subset\Lambda^q(M)$ denote all q-forms in $\Lambda^q(M)$ that vanish on U, and let $C_U\subset\Lambda^q(M)$ denote all q-forms whose supports are contained in U. Clearly, the linear subspaces C_U and Z_U in $\Lambda^q(M)$ are orthogonal with respect to both $\langle,\rangle_R$ and $\langle,\rangle_E$. It is easy to see that U is a causally q-regular neighborhood if and only if

$$C_U{}^{\perp_E}=N+_E Z_U, \tag{47}$$

where the $\langle,\rangle_E$ orthogonal sum $+_E$ is not necessarily a direct sum.

VI. GEOMETRIC CRITERIA RELATED TO THE FIELD-THEORETICAL CAUSALITY CONDITIONS

Given a complete compact space-time (M,g) of dimension$\geqslant 3$, under what conditions on M and g is (M,g) benign? In this section, we will announce some partial results that point to a resolution of this question in terms of the geometry of (M,g). Our results fall far short of providing any kind of a definitive answer. In particular, we cannot even prove that every compact manifold M [with $\chi(M)=0$] of dimension$\geqslant 3$ admits a benign Lorentz metric in $\mathrm{Lor}_c(M)$. However, we believe that this statement is true; in fact, in Sec. 6.3 of Ref. 12 we conjecture that the subset of $(q-)$ benign metrics is generic in $\mathrm{Lor}_c(M)$ whenever $\dim M\geqslant 3$. On the other hand, in dimensions higher than 2 it is, in general, exceedingly difficult to explicitly construct benign Lorentz metrics on a given compact manifold M.

As a prelude to our discussion, we will first explore the causal regularity of higher-dimensional flat Lorentz tori. For definiteness, we will focus our attention on $(q=0)$ scalar field theory on the simplest "unit" four-torus $T^4{}_{(1)}\equiv T_{(a_1,a_2,a_3,a_4)}$, where $a_1=(1,0,0,0)$, $a_2=(0,1,0,0)$, $a_3=(0,0,1,0)$, and $a_4=(0,0,0,1)$. The flat torus $T^4{}_{(1)}$ admits a large infinite-dimensional kernel N for the wave operator $\Box_Q$ (see Sec. V). [In contrast, e.g., the torus $T_{(a_1,a_2,a_3,b_4)}$ with $b_4=(0,0,0,r)$ does not admit any nonconstant solutions of the wave equation when r^2 is irrational.] Surprisingly, we will find that $T^4{}_{(1)}$ is not benign; in fact, $T^4{}_{(1)}$ is not even causally regular: we will prove that the subset $C\subset T^4{}_{(1)}$ of cr points in $T^4{}_{(1)}$ is empty.

Before describing this proof (which we will only give for the $q=0$ case), let us briefly discuss the structure of the space N of solutions to the scalar wave equation on $T^4{}_{(1)}$. Consider the initial data $\{\phi_0,\dot{\phi}_0\}$ that a solution $\phi\in N$ induces on the global spacelike hypersurface $\{t=0\}$ (with topology T^3) in $T^4{}_{(1)}$. On the covering (Minkowski) space (R^4,η), the solution ϕ that develops from the (periodically extended) data $\{\phi_0,\dot{\phi}_0\}$ is given by[29]

$$\phi(x,t)=\frac{1}{4\pi t^2}\int_{|\xi-x|=t}[t\dot{\phi}_0(\xi)+\phi_0(\xi)+\nabla\phi_0(\xi)\cdot(\xi-x)]dS_\xi, \tag{48}$$

and it coincides with the canonical (periodic) lifting [to (R^4,η)] of the solution ϕ on the quotient $T^4{}_{(1)}$. [In Eq. (48), dS_ξ denotes the area element on the sphere $|\xi-x|=t$.] The right-hand side of Eq. (48) has the correct periodicity in x by construction (since ϕ_0 and $\dot{\phi}_0$ are toroidally periodic in $\xi\in R^3$). On the other hand, the solution $\phi(x,t)$ must be one-periodic in the time coordinate t as well; according to Eq. (48) this constrains the initial data $\{\phi_0,\dot{\phi}_0\}$ so that

$$\int_{t\leqslant|\xi-x|\leqslant t+1}\frac{1}{|\xi-x|}\times\left[\Delta\phi_0+\frac{1}{|\xi-x|}[\dot{\phi}_0+(\xi-x)\cdot\nabla\dot{\phi}_0]\right]d\xi=0,\quad \forall x,\forall t>0. \tag{49}$$

It is now fairly easy to see that if N_0 denotes the linear subspace of $L^2(T^3)$ that consists of all possible initial data ϕ_0 (or $\dot{\phi}_0$) which satisfy the time-periodicity constraint Eq. (49) (i.e., which are induced from solutions $\phi\in N$), then N_0 is spanned by orthonormal basis elements of the form

$$\zeta_\alpha\equiv\zeta^{(i)}{}_{(K,L,M,q)}\equiv(2)^{3/2}\{\sin,\cos\}\left[2\pi q\frac{2KM}{d}x\right]\{\sin,\cos\}\left[2\pi q\frac{2LM}{d}y\right]\times\{\sin,\cos\}\left[2\pi q\frac{M^2-L^2-K^2}{d}z\right], \tag{50}$$

where $i=1,2,\ldots,8$ is an index for the particular choice of sines and cosines, q runs over positive integers, K,L,M run over non-negative integers, and $d=d(K,L,M)$ is the positive integer such that

$$\left(\frac{2KM}{d},\frac{2LM}{d},\frac{M^2-L^2-K^2}{d}\right)=1. \tag{51}$$

[The integers $k=2KM/d$, $l=2LM/d$, $m=(M^2-L^2-K^2)/d$, and $n=(M^2+L^2+K^2)/d$ constitute the general solution of the simple Diophantine equation

$k^2+l^2+m^2=n^2$, $(k,l,m)=1$; see Mordell.[30]]

Theorem 1: The subset $C_0 \subset T^4{}_{(1)}$ of causally zero-regular points in $T^4{}_{(1)}$ is empty.

Proof: Since $T^4{}_{(1)}$ admits a transitive group of isometries consisting of translations [Eq. (4)], it is sufficient to show the causal nonregularity of a single arbitrary point $p \in T^4{}_{(1)}$. We can choose p such that $p \in \{t=0\} \equiv T^3$. Let us assume, contrary to the statement of the theorem, that p is a c0r point; then there exists a convex normal c0r neighborhood V around p. Put $U \equiv V \cap \{t=0\}$. By construction, initial data for the scalar wave equation on $T^4{}_{(1)}$ are freely specifiable on the neighborhood $U \subset \{t=0\}$. In U, we can place a closed cube $C_Q \subset U$ with dimensions $1/Q \times 1/Q \times 1/Q$, where Q is a sufficiently large positive integer. Clearly, initial data are also freely specifiable on the cube C_Q.

The functions that are available to pose the initial data ϕ_0 and $\dot{\phi}_0$ on C_Q are given by linear combinations of the basis elements ζ_α [Eqs. (50)–(51)]. Consider ζ_α as elements of $L^2(C_Q)$ by simply restricting them to $C_Q \subset T^3$. The Hilbert space $L^2(C_Q)$ is obtained by the completion of $C^\infty(C_Q)$ with respect to the inner product

$$\langle f,g\rangle \equiv \int_0^{1/Q}\int_0^{1/Q}\int_0^{1/Q} fg\,dx\,dy\,dz, \quad f,g \in C^\infty(C_Q), \tag{52}$$

and a complete orthonormal basis for $L^2(C_Q)$ is

$$\begin{aligned} u_\alpha &\equiv u^{(i)}{}_{(K,L,M)} \\ &\equiv (2Q)^{3/2}\{\sin,\cos\}[2\pi QKx]\{\sin,\cos\}[2\pi QLy] \\ &\quad \times \{\sin,\cos\}[2\pi QMz], \\ &\langle u_\alpha,u_\beta\rangle = \delta_{\alpha\beta}, \end{aligned} \tag{53}$$

where i, K, L, and M are integers as defined before [see Eq. (50)].

The statement that initial data ϕ_0, $\dot{\phi}_0$ are freely specifiable on C_Q is equivalent to the statement that the elements $\zeta_\alpha \in L^2(C_Q)$ form a complete basis for $L^2(C_Q)$. Obviously, $\{\zeta_\alpha\}$ is a linearly independent set of vectors in $L^2(C_Q)$, but it is *not* an orthonormal (or orthogonal) sequence. We will apply the following argument to explore the completeness of the sequence $\{\zeta_\alpha\}$ in the Hilbert space $L^2(C_Q)$: Given any element $\eta \in L^2(C_Q)$ and an integer $N \geqslant 1$, we approximate η by using $\zeta_1,\zeta_2,\ldots,\zeta_N$ only; i.e., by minimizing

$$\|\eta - \sum_{k=1}^{N} c_k\zeta_k\|. \tag{54}$$

The minimum of (54) is achieved when the coefficients c_k satisfy

$$-2\langle\eta,\zeta_k\rangle + 2\sum_{l=1}^{N}\langle\zeta_k,\zeta_l\rangle c_l = 0. \tag{55}$$

With the values (55) for c_k, the minimum value achievable for $\|\eta - \Sigma_{k=1}^{N} c_k\zeta_k\|^2$ is

$$\langle\eta,\eta\rangle - \sum_{k,l=1}^{N}\langle\eta,\zeta_k\rangle G^{(N)}{}_{kl}\langle\zeta_l,\eta\rangle, \tag{56}$$

where $G^{(N)}$ is the inverse of the $N \times N$ matrix $T^{(N)}$ defined by

$$T^{(N)}{}_{ij} \equiv \langle\zeta_i,\zeta_j\rangle; \tag{57}$$

the matrix $T^{(N)}$ is invertible since ζ_α are linearly independent. [Supposing (for contradiction) that $\{\zeta_\alpha\}$ is complete, we may assume, after trimming down the sequence $\{\zeta_\alpha\}$ if necessary, that $G^{(N)}$ remains bounded as $N \to \infty$.] Now, the sequence $\{\zeta_\alpha\}$ is a complete basis for $L^2(C_Q)$ if and only if for every $\eta \in L^2(C_Q)$

$$\lim_{N\to\infty}\left[\sum_{k,l=1}^{N}\langle\eta,\zeta_k\rangle G^{(N)}{}_{kl}\langle\zeta_l,\eta\rangle\right] = \langle\eta,\eta\rangle. \tag{58}$$

We will show that there are elements $\eta \in L^2(C_Q)$ for which Eq. (58) fails to hold. This proves that $\{\zeta_\alpha\}$ is not complete in $L^2(C_Q)$, and this contradiction establishes that $p \in T^4{}_{(1)}$ cannot be a c0r point.

Let (K_0,L_0,M_0) be any triple of integers such that $K_0{}^2+L_0{}^2+L_0{}^2$ is *not* a perfect square. Then, for any integer $n \geqslant 1$, $(nK_0)^2+(nL_0)^2+(nM_0)^2$ and $(nQK_0)^2+(nQL_0)^2+(nQM_0)^2$ are not perfect squares either. Consider the infinite sequence of elements [Eq. (53)]

$$\eta_n \equiv u^{(i)}{}_{(nK_0,nL_0,nM_0)} \tag{59}$$

in $L^2(C_Q)$ for some fixed $1 \leqslant i \leqslant 8$. Clearly,

$$\langle\eta_n,\eta_n\rangle \equiv 1, \quad \forall n \geqslant 1. \tag{60}$$

On the other hand, using Eqs. (53), (52), and (50) it is straightforward to obtain the following estimate

$$|\langle\eta_n,\zeta_\alpha\rangle| < C(Q,\alpha,K_0,L_0,M_0)/\sqrt{n}, \tag{61}$$

where $C(Q,\alpha,K_0,L_0,M_0)$ is a positive quantity depending on its arguments but not on n. The crucial ingredient in establishing the estimate (61) is the fact that as $n \to \infty$, the perfect-square triple closest to $T_n \equiv (nQK_0,nQL_0,nQM_0)$ differs from T_n asymptotically as $\sim A\sqrt{n}$, where A is a constant depending on Q and (K_0,L_0,M_0) [this follows from the general solution (51) of the Diophantine problem $k^2+l^2+m^2=n^2$]. In computing $|\langle\eta_n,\zeta_\alpha\rangle|$, it is seen that for fixed α $|\langle\eta_n,\zeta_\alpha\rangle|$ is monotonically decreasing for large enough n, and that it decays asymptotically as $\sim 1/\sqrt{n}$; hence the inequality (61) is asymptotically sharp. In fact, the constant $C(Q,\alpha,K_0,L_0,M_0)$ is bounded as $|\alpha| \to \infty$, we can thus write

$$|\langle\eta_n,\zeta_\alpha\rangle| < D(Q,K_0,L_0,M_0)/\sqrt{n}, \tag{62}$$

where D is another constant independent of α and n. It is now easy to see that given any $\epsilon > 0$, there exists an integer $K_\epsilon \geqslant 1$ such that

$$|\langle\eta_n,\zeta_\alpha\rangle| < \epsilon, \quad \forall n \geqslant K_\epsilon, \forall\alpha. \tag{63}$$

Clearly [cf. Eq. (60)], when $\epsilon > 0$ is small enough, the equality (58) cannot be satisfied with those elements η_n for which $n > K_\epsilon$. This completes the proof of Theorem 1.□

The above proof shows that for any point $p \in T^4{}_{(1)}$, there are null directions at p along which the construction of plane-wave solutions with arbitrarily high frequency is prohibited by the PSC. These null directions (along which the PSC imposes an effective cutoff on the frequency of plane waves) are precisely those null vectors ξ, for which the null geodesics $\gamma(t) = \exp_p(t\xi)$ are not closed but densely fill some $\geqslant$ 2-dimensional submanifold of space-time. Furthermore, exactly the same reasoning as the above proof demonstrates that in general the flat n-torus $T^n{}_{(1)}$ is not causally q

regular for any $0 \leqslant q \leqslant n-2$ when $n \geqslant 3$. We have not studied the causal regularity of the other more general higher-dimensional flat tori, but our results here and in the following subsections suggest that it is unlikely for any of these flat tori to be benign (or causally regular) in dimensions higher than 2.

It is, in general, extremely difficult to explore the causal regularity of a higher-dimensional compact space-time (M,g) in the above manner, i.e., by studying the local-completeness properties of the space of solutions to the wave equation. One would like to have more accessible criteria involving the geometry of (M,g) that would provide necessary and sufficient conditions for (M,g) to be benign or causally regular. In the following two subsections, we will announce the results of some preliminary attempts to construct such criteria. The proofs of these results can be found, along with some speculations on benign compact space-times in higher-than-two dimensions, in Ref. 12.

A. Criteria involving the geometry of null submanifolds

Consider the scalar wave equation on the flat two-torus $T_{(1,r)}$ [Sec. IV, Eqs. (10)–(12)]. The reader might already have observed in Sec. IV that when the torus $T_{(1,r)}$ is benign (i.e., when r is rational), all of its null geodesics are closed, whereas when r is irrational [i.e., when $T_{(1,r)}$ is not benign] every null geodesic is open and densely fills the whole torus $T_{(1,r)}$. In the local coordinates $u = t - x$, $v = t + x$, $\Box\phi = 0$ reads $\phi_{,uv} = 0$; therefore, for any solution ϕ, $\phi_{,u}$ is constant along a null geodesic $\{u = \text{const}\}$, and $\phi_{,v}$ is constant along a null geodesic $\{v = \text{const}\}$. In other words, information about $d\phi$ propagates exactly (without diffraction) along the null geodesics of $T_{(1,r)}$. After this observation, it becomes easy to understand why $T_{(1,r)}$ is not benign (in fact, it does not admit any nonconstant solutions to $\Box\phi = 0$) whenever r is irrational, because for irrational r each null geodesic in $T_{(1,r)}$ winds densely around the torus and comes arbitrarily close to every point throughout the entire space-time.

In higher dimensions ($\geqslant 3$), information on solutions of the wave equation does not propagate along the null geodesics exactly, but only asymptotically in the geometric-optics limit.

Definitions: A *flat ray field* in (M,g) is a closed, null one-form $\omega \in \Lambda^1(M)$ that satisfies $d*\omega = 0$. In an even-dimensional space-time, a *flag field* is a pair (ω,h), where $\omega \in \Lambda^1(M)$ is a closed null one-form, $h \in \Lambda^{n/2-1}(M)$ is a closed spacelike $(n/2-1)$-form, and $\omega \wedge h$ spans a null plane at each point and satisfies $d*(\omega \wedge h) = 0$.

Theorem 2: Let a space-time (M,g) (not necessarily compact) admit a flag field (or a flat ray field) $\mathcal{F}$, and let $p \in M$ be a point in M. Suppose that every neighborhood of p contains a point q such that p is a limit point of the integral curve of $\mathcal{F}$ through q and q is a limit point of the integral curve of $\mathcal{F}$ through p (both in the future direction). Then p cannot be a causally regular (causally zero regular) point of (M,g).

Theorem 3: Let $\mathcal{F}$ be a flag field (flat ray field) on (M,g). Let γ be a closed (null geodesic) integral curve of $\mathcal{F}$. If $\mathcal{F}$ admits $(n-2)$ families of closed integral curves that correspond to $(n-2)$ spacelike, linearly independent variations of γ, then there exists an open neighborhood $U \subset M$ around γ such that the integral curve of $\mathcal{F}$ through every point $p \in U$ is closed, and has the same homotopy type as (i.e., is homotopic to) the integral curve γ. In particular, the following hold: (i) If all integral curves of $\mathcal{F}$ are closed, then they all have the same homotopy type (i.e., they are all homotopic). (ii) In two dimensions, either all integral curves of $\mathcal{F}$ are closed (and have the same homotopy type), or all of them are open. (iii) All integral curves of $\mathcal{F}$ are closed iff the closed null one-form ω is exact iff the null hypersurfaces orthogonal to K_ω define a smooth, $(n-1)$-dimensional foliation of M iff there is a global Maxwell field A with $dA = \omega \wedge h$ (a global solution ϕ with $d\phi = \omega$).

B. Spectral criteria

Consider the spectrum of the d'Alembertian $\Box_Q$ on functions on a flat two-torus $T_{(1,r)}$. (Recall our discussion in the last paragraph of Sec. IV.) Whenever $T_{(1,r)}$ is benign (i.e., when r is rational), the spectrum of $\Box_Q$ is as regular as that of any elliptic operator except for the infinite dimensionality of the kernel (zero eigenspace) N. In the other cases [i.e., when r is irrational and $T_{(1,r)}$ is not benign], the spectrum is not discrete and has a highly irregular structure; in general, there are many infinite-dimensional eigenspaces, and in some cases clustering of nonzero eigenvalues may occur. We will see in this subsection that such connections between the spectral geometry and causal regularity of a compact spacetime (M,g) exist more generally; in fact, as we speculate in Ref. 12 these connections are likely to probe very deep into the geometry and topology of (M,g).

Consider the following statements, where $U \subset M$ is an arbitrary open subset: (S1): For every $\psi \in V$ satisfying $\Box_Q\psi = 0$ on $M \setminus U$, $\exists$ a $\phi \in N$ that coincides with ψ on $M \setminus U$. (S2): For every $\psi \in I$ satisfying $\Box_Q\psi = 0$ on $M \setminus U$, $\exists$ a $\phi \in N$ that coincides with ψ on $M \setminus U$.

Proposition 1: The statements (S1) and (S2) are equivalent.

Theorem 4: An open neighborhood $U \subset M$ is causally q regular if and only if the statement (S1) is true. In particular, U is a cqr neighborhood if and only if the interior of $M \setminus U$ is causally q regular.

Definition: The null space (kernel of $\Box_Q$) N is called *sufficiently large* with respect to an open set $U \subset M$, if for every $\psi \in V \setminus C_U$ satisfying $\Box_Q\psi = 0$ on $M \setminus U$ there exists an element $\phi \in N$ that vanishes on U and satisfies $\langle\phi,\psi\rangle_R \neq 0$. We will simply say N is sufficiently large if every neighborhood around any point $p \in M$ contains a neighborhood U of p with respect to which N is sufficiently large.

Definition: A compact space-time (M,g) is called *spectrally q benign* if the (point) spectrum $\{\lambda_i\}$ of $\Box_Q$ satisfies the following four conditions: (i) The zero eigenspace (kernel of $\Box_Q$) N is sufficiently large. (ii) For each nonzero eigenvalue λ_i, $\dim(E_{\lambda_i}) < \infty$. (iii) The nonzero eigenvalues do not have any accumulation points in the real line. (iv) The infinite sum $\Sigma_{i \neq 0,e} |\lambda_i|^{-2}$ is convergent.

Theorem 5: A spectrally q-benign compact space-time (M,g) is q benign.

VII. COMMENTS ON QUANTUM FIELD THEORY ON A COMPACT BACKGROUND

Classical (Lorentzian) field theory on a compact space-time involves a wide variety of novel mathematical features some of which we have explored in the previous sections. In this short section we will briefly comment on the issues raised by quantization. For simplicity we restrict our attention to scalar ($q=0$) field theory on a compact background (M,g). An immediate consequence of compactness is that canonical quantization ceases to be a convenient and natural approach: A compact space-time may not admit any global spacelike hypersurfaces (Sec. II); and even when such hypersurfaces exist they may not all be homologous, i.e., there may exist a pair of hypersurfaces whose difference is not the boundary of an n-cycle in M (by contrast, all Cauchy surfaces in a globally hyperbolic space-time are homologous to each other). Consequently, it is difficult to define a natural Klein–Gordon inner product (symplectic structure) on the space of solutions N. (For a quick review of the relevant aspects of canonical quantization see the introductory section in Woodhouse.[31]) Moreover, even if one succeeds in constructing a symplectic product on N [e.g., via

$$(\phi,\psi)\equiv\int_{\Sigma}(\phi\nabla^{\alpha}\psi-\psi\nabla^{\alpha}\phi)d^{n-1}\Omega_{\alpha}$$

after singling out a particular global spacelike hypersurface Σ (which hypersurface determines each solution in N uniquely by its Cauchy data on Σ)], the standard postulates of canonical quantization (such as the commutativity of the field operators evaluated at different points of Σ) are difficult to justify because of the global causal correlations (self-consistency constraints on the data on Σ) between locally spacelike-separated points of (M,g).

We believe that the most natural approach to quantization in a compact background space-time (or in any space-time containing closed causal curves) is path integration. The path-integral approach easily makes transparent a fundamental feature of quantum field theory on compact backgrounds, namely, the existence of a uniquely determined canonical vacuum state. Consider the generating functional

$$Z[J]\equiv\int_{\{\phi\perp N\}}\mathscr{D}[\phi]\exp i\left(\int_{M}L[\phi]+\int_{M}J\phi\Omega\right),\quad(64)$$

where $\int_M L[\phi]$ is the action defined by Eq. (25)

$$L[\phi]\equiv d\phi\wedge *d\phi=g(\nabla\phi,\nabla\phi)\Omega,\quad(65)$$

and the path integral is over all fields ϕ *orthogonal* to the space of solutions N with respect to the Hilbert-space structure $\langle,\rangle_R$ on $\Lambda^0(M)$. It is easy to see that the Green's function (propagator) derived from $Z[J]$,

$$G(x,y)\equiv-2i[\delta^2\ln Z[J]/\delta J(x)\delta J(y)]_{J\equiv 0}\quad(66)$$

is precisely the Green's operator G defined in Eqs. (45)–(46). (It is instructive to compute the path integral (64) explicitly and verify this statement for the flat two-torus $T_{(1,1)}$ (Sec. IV); the result can be obtained [after expanding the field ϕ in an orthonormal basis of $\Lambda^0(M)$] via a simple regularization process. We suspect that this is a generic feature of Lorentzian path integrals on a compact background; they can be computed directly using standard regularization schemes without recourse to the Euclidean approach.) The construction Eq. (64) of the generating functional $Z[J]$ specifies a unique vacuum state for the field ϕ because the Green's function Eq. (66) is sensitive to the choice of vacuum; in other words, the canonical vacuum state is associated with a canonical choice of boundary conditions ($\phi\perp N$) for the (classical) field ϕ in the path integral (64). If (M,g) admits a global spacelike hypersurface Σ, then an alternative way to define the above vacuum is to introduce the "vacuum-state wave functional" $\Psi_0:C^{\infty}(\Sigma)\to R$ via the path integral

$$\Psi_0[h]\equiv\int_{\{\phi|_{\Sigma}=h,\phi\perp N\}}\mathscr{D}[\phi]\times\exp\left(i\int_{M}L[\phi]\right)\quad[h\in C^{\infty}(\Sigma)],\quad(67)$$

where the integration is over all fields $\phi\perp N$ that coincide with h on Σ. It is an interesting challenge to give a canonical-quantization interpretation for the vacuum defined by Eq. (67) [i.e., an interpretation based on a specific choice of polarization (mode decomposition) in the classical phase space of solutions N]. (Quantum field theory in curved space-time is reformulated in Ref. 32 in a way that may prove particularly useful in the above context.)

If our Conjecture 5 in Sec. 6.3 of Ref. 12 is false and nonbenign compact space-times are the rule rather than the exception, would this necessarily render the study of compact space-times an irrelevant mathematical exercise? It appears conceivable that from the viewpoint of quantum field theory nonbenign space-times may be as interesting as benign ones. For example, a nonbenign two-torus $T_{(1,r)}$ with r irrational (Sec. IV) features a *finite-dimensional* phase space of classical solutions, and it is easy to construct other examples of compact space-times (consider $S^2\times S^1$) whose phase spaces of solutions (to $\Box\phi=0$) have arbitrary and finite dimensions. On these compact backgrounds the vacuum expectation value of the stress-energy tensor for the field ϕ is manifestly finite. On the other hand, near any point in any of these space-times there exist as many *local* modes of solutions to $\Box\phi=0$ as in Minkowski space; most of these local modes do not extend globally to become elements of the phase space N. Since the standard argument for stress-tensor renormalization is the "subtraction" of the contribution of these local high-frequency modes, it is not entirely obvious how the avoidance of renormalization can be justified when N is finite dimensional. Similar issues of principle are also raised for those nonbenign space-times where the phase space N is not finite dimensional but excludes all (or almost all) local modes with high enough frequency (a good example is the flat four-torus $T^4_{(1)}$ discussed in Sec. VI).

ACKNOWLEDGMENTS

The author would like to thank the International Center for Theoretical Physics for its hospitality, and Danny Birmingham, Matthias Blau, George Ellis, Robert Geroch, Caren Marzban, Dennis Sciama, and Robert Wald for useful discussions at various stages of this work.

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

[2] R. Penrose, "Singularities and Time Asymmetry," in *General Relativity: An Einstein Centenary Survey*, edited by S. W. Hawking, and W. Israel, (Cambridge U. P., Cambridge, 1979).

[3] C. J. S. Clarke, G. F. R. Ellis and F. J. Tipler, "Singularities and Horizons: A Review Article," in *Einstein Centenary Volume*, edited by P. Bergmann and A. Held, (Plenum, New York, 1979).
[4] M. S. Morris, K. S. Thorne, and U. Yurtsever, Phys. Rev. Lett. **61**, 1446 (1988).
[5] T. A. Roman, Phys. Rev. D **37**, 546 (1988).
[6] J. Friedman, G. Klinkhammer, F. Echeverria, M. S. Morris, I. D. Novikov, K. S. Thorne, and U. Yurtsever, "The Cauchy Problem in Spacetimes With Closed Timelike Curves," Phys. Rev. D, submitted. (Also available as Caltech preprint GRP-000.)
[7] I. M. Singer, "Eigenvalues of the Laplacian and Invariants of Manifolds," in *Proceedings of the International Congress of Mathematicians*, Vancouver, 1974, pp. 187–200.
[8] D. Freed, and K. Uhlenbeck, *Instantons and Four-Manifolds* (Springer-Verlag, Berlin, 1984).
[9] M. F. Atiyah, "New Invariants of Three and Four-Dimensional Manifolds," in *The Symposium on the Mathematical Heritage of Hermann Weyl*, edited by R. Wells *et al.* (University of North Carolina, Chapel Hill, 1987).
[10] E. Witten, Commun. Math. Phys. **117**, 353 (1988).
[11] A. Meister, "Complex-Manifold Methods in Quantum Field Theory in Curved Spacetime," University of Bern Theoretical Physics preprint BUTP-89/5, March 1989.
[12] U. Yurtsever, "Test Fields on Compact Spacetimes: Problems, Some Partial Results, and Speculations," ICTP preprint IC-89-308.
[13] L. Markus, Ann. Math. **62**, 411 (1955).
[14] M. W. Hirsch, *Differential Topology* (Springer-Verlag, Berlin, 1976).
[15] E. H. Spanier, *Algebraic Topology* (Springer-Verlag, Berlin, 1983).
[16] K. K. Lee, Gen. Relat. Gravit. **6**, 239 (1975).
[17] K. T. Chen, Proc. Am. Math. Soc. **26**, 196 (1970).
[18] R. P. A. C. Newman, Gen. Relat. Gravit. **18**, 1181 (1986).
[19] R. P. Geroch, J. Math. Phys. **9**, 1793 (1968).
[20] G. S. Whiston, J. Phys. A **11**, 1203 (1978).
[21] C. W. Misner and A. H. Taub, Zh. Eksp. Teor. Fiz. **55**, 233 (1968) [Sov. Phys. JETP **28**, 122 (1969)].
[22] C. W. Misner, "Taub–Nut Spacetime as a Counterexample to Almost Anything," in *Relativity Theory and Astrophysics*, edited by J. Ehlers (American Mathematical Society, Providence, RI, 1967).
[23] D. Fried, J. Diff. Geom. **26**, 385 (1987).
[24] W. M. Goldman and Y. Kamishima, J. Diff. Geom. **19**, 233 (1984).
[25] F. W. Warner, *Foundations of Differentiable Manifolds and Lie Groups* (Springer-Verlag, Berlin, 1983).
[26] H. Farkas and I. Kra, *Riemann Surfaces* (Springer-Verlag, Berlin, 1980).
[27] T. Dray, R. Kulkarni, and J. Samuel, J. Math. Phys. **30**, 1306 (1989).
[28] R. D. Richtmyer, *Principles of Advanced Mathematical Physics Volume I* (Springer-Verlag, Berlin, 1978).
[29] F. John, *Partial Differential Equations* (Springer-Verlag, Berlin, 1982).
[30] L. Mordell, *Diophantine Equations* (Academic, London, 1969).
[31] N. Woodhouse, Proc. R. Soc. London Ser. A **378**, 119 (1981).
[32] B. Kay and R. Wald, "Vacuum States in Spacetimes with Bifurcate Killing Horizons," Enrico Fermi Institute Preprint, 1989 (to appear in Phys. Rep.).

New link polynomial obtained from octet representation of quantum sl (3) enveloping algebra

Zhong-Qi Ma[a)]
CCAST (World Laboratory) P.O. Box 8730, Beijing, People's Republic of China and International Centre for Theoretical Physics, Trieste, Italy

(Received 3 August 1989; accepted for publication 5 July 1990)

The quantum Clebsch–Gordan coefficients for the coproduct 8×8 of the quantum sl(3) enveloping algebra are computed. In the decomposition of the coproduct 8×8, there are two octet representations that are identified by the symmetry in changing the order of the factor octet representations. The corresponding $\check{R}_q$ matrix and a new link polynomial are obtained.

I. INTRODUCTION

The quantum Yang–Baxter equation (QYBE)[1] plays a crucial role in the completely integrable system of two dimension,[2] and has an intimate relation to the conformal field theory.[3] From the solution R of QYBE, a representation of the braid group and a link polynomial can be built up.[4,5] QYBE is a highly nonlinear equation and is difficult to solve. When the quantum parameter q tends to unity, it becomes the classical one (CYBE), which appears to be a commutation relation. The solutions of CYBE were classified.[6] Quantizing the trigonometric solution of CYBE based on a simple Lie algebra,[7] we otain the solution $\check{R}_q$ of QYBE, which can be expressed by the quantum Clebsch–Gordan (qCG) coefficients of the quantum enveloping algebra.[8,9] The $\check{R}_q$ matrices for any irreducible representations (IR) of the quantum sl(2) enveloping algebra [q-sl(2)] were discussed in great detail.[8,10] Those for the fundamental representations of some quantum classical Lie algebras were discussed, and proved to satisfy the Hecke algebra.[11,12] In our previous paper,[13] we computed the $\check{R}_q$ matrix and the corresponding link polynomial for the coproduct 6×6 of the quantum sl(3) enveloping algebra [q-sl(3)]. In all these cases, there is no multiplicity in the decomposition of a coproduct.

The decomposition of the coproduct 8×8 of q-sl(3) is the simplest but very useful example where there is multiplicity in the decomposition, namely, there are two octet representations appearing in the coproduct 8×8 of q-sl(3). In the SU(3) theory, those two octet representations are identified by the symmetry for changing the order of the factor octet representations.[14] This symmetry can be generalized to the case of the quantum enveloping algebras in discussing the decomposition problem with multiplicity. In this paper we discuss this typical example in detail. We will compute the q-CG coefficients and the $\check{R}_q$ matrix, and then, obtain a new link polynomial.

In this paper we use the same notations as those used in the previous paper.[13] The plan of this paper is as follows. In Sec. II, we describe the octet representaiton of q-sl(3) briefly. In Sec. III, the qCG coefficients of the coproduct 8×8 of q-sl(3) are given. We review the general method of computing the $\check{R}_q$ matrix, which is the solution of QYBE, and give the calculation results based on the octet representation of q-sl(3). At last, a new link polynomial is built up in terms of the representation of the braid group based on this $\check{R}_q$ matrix in Sec. V.

II. OCTET REPRESENTATION OF q-sl(3)

We denote an IR of q-sl(3) by its Young tableau $[\lambda_1,\lambda_2]$ or, briefly, by its dimension N. The enumeration of the states of the octet representation is given in Fig. 1(a). Using the method given in the previous paper,[13] we obtain the octet representation matrices from the decomposition $[2,0]\times[1,0]=[3,0]+[2,1]$ $(6\times3=10+8)$:

$$D_q^8(k_1)=\mathrm{diag}(q^{1/2},q^{-1/2},q,1,1,q^{-1},q^{1/2},q^{-1/2}),$$
$$D_q^8(k_2)=\mathrm{diag}(q^{1/2},q,q^{-1/2},1,1,q^{1/2},q^{-1},q^{-1/2}),$$
$$D_q^8(e_1)_{12}=D_q^8(e_1)_{78}=D_q^8(e_2)_{13}=D_q^8(e_2)_{68}=1,$$
$$D_q^8(e_1)_{34}=D_q^8(e_1)_{46}=[2]^{1/2},$$
$$D_q^8(e_2)_{24}=D_q^8(e_2)_{47}=[2]^{-1/2},$$
$$D_q^8(e_2)_{25}=D_q^8(e_2)_{57}=([3]/[2])^{1/2}. \tag{1a}$$

The rest of the matrix elements are vanishing, and the matrices of f_a are the transpose of those of e_a;

$$D_q^8(f_a)=\widetilde{D_q^8(e_a)},\quad a=1,2. \tag{1b}$$

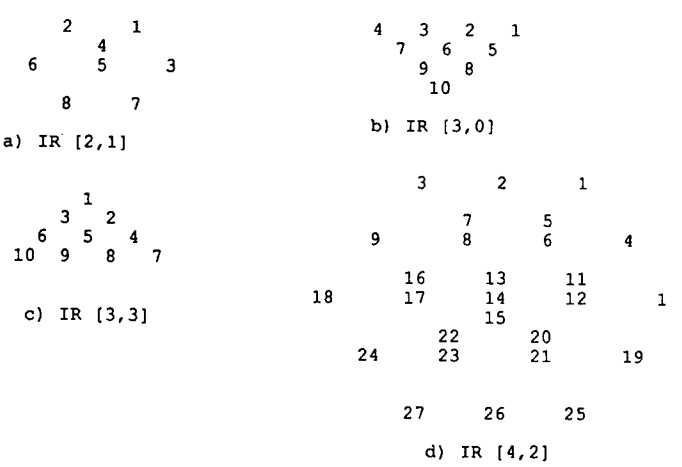

FIG. 1. The enumerations of the states of IR's of q-sl(3).

[a)] On leave from the Institute of High Energy Physics, P.O. Box 918(4), Beijing 100039, People's Republic of China.

 0022-2488/90/123079-06$03.00

III. qCG COEFFICIENTS FOR THE COPRODUCT 8×8

The coproduct 8×8 is defined as

$$D_q^{88}(k_a) = D_q^8(k_a) \times D_q^8(k_a),$$
$$D_q^{88}(e_a) = D_q^8(e_a) \times D_q^8(k_a^{-1}) + D_q^8(k_a) \times D_q^8(e_a) \tag{2}$$

and that replaced e_a by f_a.

The representation D_q^{88} is a reducible one and can be reduced by an orthogonal transformation C_q, called the qCG matrix,

$$C_q^{-1} D_q^{88}(I) C_q = \bigoplus_N D_q^N(I), \tag{3}$$

where $N = 27, 10, 10^*, 8_S, 8_A$ and 1, i.e., [4,2], [3,0], [3,3], $[2,1]_S$, $[2,1]_A$, and [0,0]. Here C_q is a 64×64 matrix with the row indices m_1, m_2 and the column indices NM. Sometimes, we use the submatrix $(C_q)_N$ which is a 64×N matrix with the column indices M. In the classical case, the CG matrix has the symmetry property for exchanging the row indices m_1 and m_2. Because of the coproduct (2), the quantum (C_q) matrix has the following symmetry property:

$$(C_q)_{m_1 m_2 NM} = \xi_N (C_{q^{-1}})_{m_2 m_1 NM}, \tag{4a}$$

where

$$\xi_{27} = \xi_{8_S} = \xi_1 = 1,$$
$$\xi_{10} = \xi_{10^*} = \xi_{8_A} = -1. \tag{4b}$$

The subscript of the octet representation just denotes this symmetry.

Through a straightforward but tedious calculation, we obtain the representation matrices and the qCG coefficients. Owing to our conventions,[13] what is only needed to be listed is the nonvanishing matrix elements of e_2 (see Table I). The enumerations of the states of IR's are given in Fig. 1. The rows of the qCG matrix 8×8 in q-sl(3) are ordered by m_1 and m_2, which both go from 1 to 8, and the columns are denoted by (27,M), (10,M), (10*,M), (8_S,M), (8_A,M), and (1,M) for the different IR's, respectively. The qCG matrix is a block matrix, and some submatrices are equal to each other. The nonvanishing matrix elements are listed in Table II. The equal submatrices are listed in the same table and distingushed by (a), (b), and so on.

In addition to those listed in Table II, there is one 10×10 submatrix in qCG matrix that combines the states $|m_1\rangle|m_2\rangle$ of the product spaces into the states $|NM\rangle$ of the irreducible representation N as follows:

$$|27,13\rangle = [2]^{1/2}([4][3])^{-1/2}\{q^2|3\rangle|6\rangle + [2]|4\rangle|4\rangle + q^{-2}|6\rangle|3\rangle\},$$
$$|27,14\rangle = \{[5][4][2]^3\}^{1/2}\{q^2[4]|1\rangle|8\rangle + q[4]|2\rangle|7\rangle + q^{-2} - q^2)|3\rangle|6\rangle$$
$$+ ([4] - 2[2])|4\rangle|4\rangle + [4][3]^{1/2}|4\rangle|5\rangle + [4][3]^{1/2}|5\rangle|4\rangle$$
$$+ (q^2 - q^{-2})|6\rangle|3\rangle + q^{-1}[4]|7\rangle|2\rangle + q^{-2}|[4]\rangle|1\rangle\},$$
$$|27,15\rangle = \{[5][4][3][2]\}^{-1/2}\{-q^2[3]|1\rangle|8\rangle + q^3[3]|2\rangle|7\rangle - q^{-1}|3\rangle|6\rangle$$
$$+ |4\rangle|4\rangle + [3]^2|5\rangle|5\rangle - q|6\rangle|3\rangle + q^{-3}[3]|7\rangle|2\rangle - q^{-2}[3]|8\rangle|1\rangle\},$$
$$|10,6\rangle = [2]^{-1}[3]^{-1/2}\{q|1\rangle|8\rangle + |2\rangle|7\rangle + |3\rangle|6\rangle + (q^{-1} - q)|4\rangle|4\rangle$$
$$- q[3]^{1/2}|4\rangle|5\rangle + q^{-1}[3]^{1/2}|5\rangle|4\rangle - |6\rangle|3\rangle - |7\rangle|2\rangle - q^{-1}|8\rangle|1\rangle\},$$
$$|10^*,5\rangle = [2]^{-1}[3]^{-1/2}\{q|1\rangle|8\rangle + |2\rangle|7\rangle + |3\rangle|6\rangle + (q^{-1} - q)|4\rangle|4\rangle$$
$$+ q^{-1}[3]^{1/2}|4\rangle|5\rangle - q[3]^{1/2}|5\rangle|4\rangle - |6\rangle|3\rangle - |7\rangle|2\rangle - q^{-1}|8\rangle|1\rangle,$$
$$|8_S,4\rangle = [3]^{1/2}([2] - 1)^{-1}\{2q[2]([3] + [2])\}^{-1/2}\{|1\rangle|8\rangle + q^{-1}|2\rangle|7\rangle$$
$$+ (1 - q)|3\rangle|6\rangle + q^{-1}(1 - q - q^2 + q^3)|4\rangle|4\rangle - q^{-1}(1 + q^3)[3]^{-1/2}$$
$$\times(|4\rangle|5\rangle + |5\rangle|4\rangle) - (1 - q)|6\rangle|3\rangle + q^2|7\rangle|2\rangle + q|8\rangle|1\rangle\},$$
$$|8_S,5\rangle = \{2q[2]([3] + [2])\}^{1/2}\{|1\rangle|8\rangle - q|2\rangle|7\rangle + q^{-2}(1 + q^3)([2] - 1)^{-1}|3\rangle|6\rangle$$
$$- q^{-1}(1 + q^3)([2] - 1)^{-1}|4\rangle|4\rangle + q^{-2}(1 - q + q^2 + q^3 - q^4 + q^5)([2] - 1)^{-1}|5\rangle|5\rangle$$
$$+ (1 + q^3)([2] - 1)^{-1}|6\rangle|3\rangle - |7\rangle|2\rangle + q|8\rangle|1\rangle\},$$
$$|8_A,4\rangle = [3]^{1/2}([2] - 1)^{-1}\{2q[2]([3] - [2])\}^{-1/2}\{-|1\rangle|8\rangle - q^{-1}|2\rangle|7\rangle$$
$$+ (1 - q)|3\rangle|6\rangle + q^{-1}(1 + q - q^2 - q^3)|4\rangle|4\rangle - q^{-1}(1 - q^3)[3]^{-1/2}$$
$$\times(|4\rangle|5\rangle + |5\rangle|4\rangle) - (1 + q)|6\rangle|3\rangle + q^2|7\rangle|2\rangle + q|8\rangle|1\rangle\},$$
$$|8_A,5\rangle = \{2q[2]([3] - [2])\}^{1/2}\{|1\rangle|8\rangle - q|2\rangle|7\rangle + q^{-2})(1 - q^3)([2] + 1)^{-1}|3\rangle|6\rangle$$
$$- q^{-1}(1 - q^3)([2] + 1)^{-1}|4\rangle|4\rangle - q^{-2}(1 + q + q^2 - q^3 - q^4 - q^5)([2] + 1)^{-1}|5\rangle|5\rangle$$
$$+ (1 - q^3)([2] + 1)^{-1}|6\rangle|3\rangle + |7\rangle|2\rangle - q|8\rangle|1\rangle\},$$
$$|1,1\rangle = ([4][2])^{-1/2}\{q^{-2}|1\rangle|8\rangle - q^{-1}|2\rangle|7\rangle - q^{-1}|3\rangle|6\rangle + |4\rangle|4\rangle + |5\rangle|5\rangle$$
$$- q|6\rangle|3\rangle - q|7\rangle|2\rangle + q^2|8\rangle|1\rangle\}.$$

IV. $\check{R}_q$ MATRIX

In the previous paper,[13,8] we gave the definition for the $\check{R}_q$ matrix. From the definition, $\check{R}_q$ matrix can be expressed by the qCG coefficients

$$\check{R}_q = \sum_N \xi_N \, q^{\eta(N)} (C_q)_N (\tilde{C}_q)_N, \tag{5}$$

where $N = 27, 10, 10^*, 8_S, 8_A$, and 1, ξ_N is given in (4b), and

$$\eta(N) = 2C_2(8) - C_2(N).$$

Removing a factor q^{-2} for normalizing, we replace η_N, by η'_N in (5),

$$\eta'_{27} = 0, \quad \eta'_{10} = \eta'_{10^*} = 2, \quad \eta'_{8_S} = \eta'_{8_A} = 5, \eta'_1 = 8. \tag{6}$$

The $\check{R}_q$ matrix is a block matrix and some submatrices are equal to each other. The calculation results are listed in Table III. The equal submatrices are listed in the same table, and distingushed by (a), (b), and so on.

There is a 10×10 submatrix in $\check{R}_q$. Notice that $\check{R}_q$ is a symmetric matrix. The nonvanishing elements of the 10×10 submatrix of $\check{R}_q$ are listed as follows:

$$(\check{R}_q)_{44,44} = (\check{R}_q)_{45,54} = (\check{R}_q)_{55,55} = q^2,$$
$$(\check{R}_q)_{18,81} = (\check{R}_q)_{27,72} = (\check{R}_q)_{36,63} = q^4,$$
$$(\check{R}_q)_{27,81} = (\check{R}_q)_{36,81} = q^3 - q^5,$$
$$(\check{R}_q)_{44,63} = q - q^5,$$
$$(\check{R}_q)_{44,72} = (q^2 - q^4)/[2],$$
$$(\check{R}_q)_{44,81} = (q^7 - 2q^3 + q)/[2],$$
$$(\check{R}_q)_{45,72} = (\check{R}_q)_{54,72} = (q^2 - q^4)[3]^{1/2}/[2],$$
$$(\check{R}_q)_{45,81} = (\check{R}_q)_{54,81} = (q - q^3)[3]^{1/2}/[2],$$
$$(\check{R}_q)_{55,72} = -q^{-1}(\check{R}_q)_{55,81} = (1 - q^6)/[2],$$
$$(\check{R}_q)_{63,63} = (\check{R}_q)_{72,72} = q^6 - q^4 - q^2 + 1,$$
$$(\check{R}_q)_{63,72} = q^6 - q^4,$$
$$(\check{R}_q)_{63,81} = (\check{R}_q)_{72,81} = -q^7 - 2q^5 - q^3,$$
$$(\check{R}_q)_{81,81} = q^8 - 2q^6 + 2q^4 - 2q^2 + 1.$$

Define

$$X_{m_1 M M' m'_1} = \sum_{m''_1 m'_2 m'_3 m_2 m_3} (C_q)_{m_2 m_3 N M} (\check{R}_q)_{m_1 m_2 m'_2 m''_1} \times (\check{R}_q)_{m''_1 m_3 m'_3 m'_1} (C_q)_{m'_2 m'_3 N M'}.$$

It is easy to show that the X matrix satisfies the definition of $\check{R}_q^{N8}$ matrix, namely, $\check{R}_q$ matrices satisfy the pentagonal relation

$$\sum_{m''_1 m'_2 m'_3} (\check{R}_q)_{m_1 m_2 m'_2 m''_1} (\check{R}_q)_{m''_1 m_3 m'_3 m'_1} (C_q)_{m'_2 m'_3 N M} = \sum_{M'} (C_q)_{m_2 m_3 N M'} (\check{R}_q^{N8})_{m_1 M' M m'_1}. \tag{7}$$

It is generally correct that the $\check{R}_q$ matrix defined in the previous paper[13] satisfies the pentagonal relation, and then, satisfies the QYBE without the spectral parameter.[9]

TABLE I. The nonvanishing matrix elements of e_2 in IR's [3,0], [3,3] and [4,2].

Row			2		3		6		4		9	7
Column			5		6		8		7		10	9
$D_q^{10}(e_2)$			1		$[2]^{1/2}$				$[3]^{1/2}$			$[2]$
Row			6		3		5		1		4	2
Column			9		5		8		2		7	4
$D_q^{10^*}(e_2)$			1		$[2]^{1/2}$				$[3]^{1/2}$			$[2]$
Row	1	4	18	24	2	22	2	23	5	16		
Column	4	10	24	27	5	26	6	26	11	22		
$D_q^{27}(e_2)$		$[2]^{1/2}$			$\frac{[2]}{[3]^{1/2}}$		$\left\{\frac{[5]}{[3]}\right\}^{1/2}$		$\left\{\frac{[3][2]}{[4]}\right\}^{1/2}$			
Row	5	17	6	17	11	9	12	9				
Column	12	22	12	23	19	16	19	17				
$D_q^{27}(e_2)$	$\left\{\frac{[5][2]}{[4][3]}\right\}^{1/2}$		$\left\{\frac{[4][2]}{[3]}\right\}^{1/2}$		$\left\{\frac{[2]}{[4]}\right\}^{1/2}$		$\left\{\frac{[5][2]}{[4]}\right\}^{1/2}$					
Row	3	20	3	21	7	13	7	14				
Column	7	25	8	25	13	20	14	20				
$D_q^{27}(e_2)$	$\left\{\frac{[2]}{[3]}\right\}^{1/2}$		$\left\{\frac{[5][2]}{[3]}\right\}^{1/2}$		$\frac{[2]}{[4]^{1/2}}$		$[2]\left\{\frac{[5]}{[4][3]}\right\}^{1/2}$					
Row	8	14	8	15								
Column	14	21	15	21								
$D_q^{27}(e_2)$	$\left\{\frac{[4]}{[3]}\right\}^{1/2}$		$[4]^{1/2}$									

TABLE II. Nonvanishing matrix elements of qCG for coproduct 8×8 in q-sl (3).

Row	Column	(a) (d)	(27,1) (27,18)	(b) (e)	(27,3) (27,25)	(c) (f)	(27,10) (27,27)
(a) 11 (b) 22 (c) 33 (d) 66 (e) 77 (f) 88					1		

Row	Column	(a) (c) (e)	(27,2) (27,9) (27,24)	(b) (d) (f)	(27,4) (27,19) (27,26)	(a) (c) (e)	(10*,1) (10,4) (10*,10)	(b) (d) (f)	(10,1) (10*,7) (10,10)
(a) 12 (b) 13 (c) 26 (d) 37 (e) 68 (f) 78			$\left\{\frac{q}{[2]}\right\}^{1/2}$				$(q\,[2])^{-1/2}$		
(a) 21 (b) 31 (c) 62 (d) 73 (e) 86 (f) 87			$(q[2])^{-1/2}$				$-\left\{\frac{q}{[2]}\right\}^{1/2}$		

Row	Column	(a) (b)	(27,5) (27,22)	(27,6) (27,23)	(10,2) (10*,9)	(10*,2) (10,9)	(8_s,1) (8_s,8)	(8_A,1) (8_A,8)
(a) 14 (b) 48			$q[3]^{-1/2}$	$\frac{-q}{[2]([5][3])^{1/2}}$	$[3]^{-1/2}$	$[3]^{-1/2}[2]^{-1}$	$N_S q^{-3/2}[3]^{1/2}$	$N_A q^{-3/2}[3]^{1/2}$
(a) 15 (b) 58			0	$\frac{q[3]}{[2][5]^{1/2}}$	0	$[2]^{-1}$	$N_S q^{-3/2}([2]-1)$	$-N_A q^{-3/2}([2]+1)$
(a) 23 (b) 67			$\frac{q^{-1/2}}{([3][2])^{1/2}}$	$q^{5/2}([5][3][2])^{-1/2}$	$\frac{q^{-3/2}}{([3][2])^{1/2}}$	$\frac{-q^{3/2}}{([3][2])^{1/2}}$	$-N_S([3][2])^{1/2}$	$-N_A([3][2])^{1/2}$
(a) 32 (b) 76			$\frac{q^{1/2}}{([3][2])^{1/2}}$	$\frac{q^{-5/2}}{([5][3][2])^{1/2}}$	$\frac{-q^{3/2}}{([3][2])^{1/2}}$	$\frac{q^{-3/2}}{([3][2])^{1/2}}$	$-N_S([3][2])^{1/2}$	$N_A([3][2])^{1/2}$
(a) 41 (b) 84			$q^{-1}[3]^{-1/2}$	$\frac{-q^{-1}}{[2]([5][3])^{1/2}}$	$-[3]^{-1/2}$	$-[3]^{-1/2}[2]^{-1}$	$N_S q^{3/2}[3]^{1/2}$	$-N_A q^{3/2}[3]^{1/2}$
(a) 51 (b) 85			0	$\frac{q^{-1}[3]}{[2][5]^{1/2}}$	0	$-[2]^{-1}$	$N_S q^{3/2}([2]-1)$	$N_A q^{3/2}([2]+1)$

Row	Column	(a) (b)	(27,7) (27,20)	(27,8) (27,21)	(10,3) (10*,8)	(10*,3) (10,8)	(8_S,2) (8_S,7)	(8_A,2) (8_A,7)
(a) 16 (b) 38			$\frac{q^{3/2}}{([3][2])^{1/2}}$	$\frac{-q^{3/2}}{([5][3][2])^{1/2}}$	$\frac{q^{1/2}}{([3][2])^{1/2}}$	$\frac{q^{1/2}}{([3][2])^{1/2}}$	$N_S q^{-1}([3][2])^{1/2}$	$N_A q^{-1}([3][2])^{1/2}$
(a) 24 (b) 47			$[3]^{-1/2}$	$\frac{q^3}{[2]([5][3])^{1/2}}$	$q^{-1}[3]^{-1/2}$	$\frac{-q^2}{[2][3]^{1/2}}$	$-N_S q^{1/2}[3]^{1/2}$	$-N_A q^{1/2}[3]^{1/2}$
(a) 25 (b) 57			0	$\frac{q[3]}{[2][5]^{1/2}}$	0	$[2]^{-1}$	$-N_S q^{-3/2}([2]-1)$	$N_A q^{-3/2}([2]+1)$
(a) 42 (b) 74			$[3]^{-1/2}$	$\frac{q^{-3}}{[2]([5][3])^{1/2}}$	$-q[3]^{-1/2}$	$\frac{q^{-2}}{[2][3]^{1/2}}$	$-N_S q^{-1/2}[3]^{1/2}$	$N_A q^{-1/2}[3]^{1/2}$
(a) 52 (b) 75			0	$\frac{q^{-1}[3]}{[2][5]^{1/2}}$	0	$-[2]^{-1}$	$N_S q^{3/2}([2]-1)$	$N_A q^{3/2}([2]+1)$
(a) 61 (b) 83			$\frac{q^{-3/2}}{([3][2])^{1/2}}$	$\frac{-q^{-3/2}}{([5][3][2])^{1/2}}$	$\frac{-q^{-1/2}}{([3][2])^{1/2}}$	$\frac{-q^{-1/2}}{([3][2])^{1/2}}$	$-N_S q([3][2])^{1/2}$	$-N_A q([3][2])^{1/2}$

Row	Column	(a) (b)	(27,11) (27,16)	(27,12) (27,17)	(10,5) (10*,6)	(10*,4) (10,7)	(8_S,3) (8_S,6)	(8_A,3) (8_A,6)
(a) 17 (b) 28			0	$\frac{q^{3/2}}{[2]}\left(\frac{[4]}{[5]}\right)^{1/2}$	$\frac{q^{1/2}}{([3][2])^{1/2}}$	$\frac{q^{1/2}}{([3][2])^{1/2}}$	$N_S q^{-1}([3][2])^{1/2}$	$-N_A q^{-1}([3][2])^{1/2}$
(a) 34 (b) 46			$q\left(\frac{[2]}{[4]}\right)^{1/2}$	$\frac{q^{-2}-1}{([5][4][2])^{1/2}}$	$\frac{q^{-1}}{[2][3]^{1/2}}$	$\frac{q^{-1}}{[2][3]^{1/2}}$	$-N_S(q^{-1/2}-q^{-3/2})[3]^{1/2}$	$N_A(q^{-1/2}+q^{-3/2})[3]^{1/2}$

TABLE II. (*Continued.*)

(a) 35 (b) 56	0	$\frac{1}{[2]}\left(\frac{[4][3]}{[5][2]}\right)^{1/2}$	$-q[2]^{-1}$	$q^{-1}[2]^{-1}$	$-N_S(q^{3/2}+q^{-3/2})$	$N_A(q^{3/2}-q^{-3/2})$
(a) 43 (b) 64	$q^{-1}\left(\frac{[2]}{[4]}\right)^{1/2}$	$\frac{q^2-1}{([5][4][2])^{1/2}}$	$\frac{-q}{[2][3]^{1/2}}$	$\frac{-q}{[2][3]^{1/2}}$	$N_S(q^{3/2}-q^{1/2})[3]^{1/2}$	$-N_A(q^{3/2}+q^{1/2})[3]^{1/2}$
(a) 53 (b) 65	0	$\frac{1}{[2]}\left(\frac{[4][3]}{[5][2]}\right)^{1/2}$	$q^{-1}[2]^{-1}$	$-q[2]^{-1}$	$-N_S(q^{3/2}+q^{-3/2})$	$N_A(q^{3/2}-q^{-3/2})$
(a) 71 (b) 82	0	$\frac{q^{-3/2}}{[2]}\left(\frac{[4]}{[5]}\right)^{1/2}$	$\frac{-q^{-1/2}}{([3][2])^{1/2}}$	$\frac{-q^{-1/2}}{([3][2])^{1/2}}$	$N_S q([3][2])^{1/2}$	$N_A q[3][2])^{1/2}$

$N_S=\{2[2]([2]-1)([4]+1)\}^{-1/2}$, $N_A=\{2[2]([2]+1)([4]-1)\}^{-1/2}$.

TABLE III. The matrix based on IR [2,1] of q-sl (3).

Row	Column (a) 11 (d) 66	(b) 22 (e) 77	(c) 33 (f) 88
(a) 11 (b) 22 (c) 33 (d) 66 (e) 77 (f) 88	1		

Row	Column (a) 12 (c) 26 (e) 68	(b) 13 (d) 37 (f) 78	(a) 21 (c) 62 (e) 86	(b) 31 (d) 73 (f) 87
(a) 12 (b) 13 (c) 26 (d) 37 (e) 68 (f) 78	0		q	
(a) 21 (b) 31 (c) 62 (d) 73 (e) 86 (f) 87	q		$1-q^2$	

Row	Column (a) 14 (b) 48	(a) 15 (b) 58	(a) 23 (b) 67	(a) 32 (b) 76	(a) 41 (b) 84	(a) 51 (b) 85
(a) 14 (b) 48	0	0	0	0	q^2	0
(a) 15 (b) 58	0	0	0	0	0	q^2
(a) 23 (b) 67	0	0	0	q^3	$(q^{3/2}-q^{7/2})[2]^{1/2}$	0
(a) 32 (b) 76	0	0	q^3	0	$(q^{3/2}-q^{7/2})[2]^{-1/2}$	$(q^{3/2}-q^{7/2})\left(\frac{[3]}{[2]}\right)^{1/2}$
(a) 41 (b) 84	q^2	0	$(q^{3/2}-q^{7/2})[2]^{1/2}$	$(q^{3/2}-q^{7/2})[2]^{-1/2}$	$(q^5-2q^1+q^{-1})[2]^{-1}$	$(q^5-q^3)\frac{[3]^{1/2}}{[2]}$
(a) 51 (b) 85	0	q^2	0	$(q^{3/2}-q^{7/2})\left(\frac{[3]}{[2]}\right)^{1/2}$	$(q^5-q^3)\frac{[3]^{1/2}}{[2]}$	$\frac{q^{-1}-q^5}{[2]}$

Row	Column (a) 16 (b) 38	(a) 24 (b) 47	(a) 25 (b) 57	(a) 42 (b) 74	(a) 52 (b) 75	(a) 61 (b) 83
(a) 16 (b) 38	0	0	0	0	0	q^3
(a) 24 (b) 47	0	0	0	q^2	0	$(q^{3/2}-q^{7/2})[2]^{1/2}$
(a) 25 (b) 57	0	0	0	0	q^2	0
(a) 42 (b) 74	0	q^2	0	$(q-q^3)[2]^{-1}$	$(q-q^3)\frac{[3]^{1/2}}{[2]}$	$(q^{1/2}-q^{5/2})[2]^{-1/2}$

TABLE III. (*Continued.*)

(a) 52 (b) 75	0	0	q^2	$(q-q^3)\frac{[3]^{1/2}}{[2]}$	$(q^{-1}-q^5)[2]^{-1}$	$(q^{9/2}-q^{5/2})\left(\frac{[3]}{[2]}\right)^{1/2}$
(a) 61 (b) 83	q^3	$(q^{3/2}-q^{7/2})[2]^{1/2}$	0	$(q^{1/2}-q^{5/2})[2]^{-1/2}$	$(q^{9/2}-q^{5/2})\left(\frac{[3]}{[2]}\right)^{1/2}$	q^4-2q^2+1

Row \ Column	(a) 17 (b) 28	(a) 34 (b) 46	(a) 35 (b) 56	(a) 43 (b) 64	(a) 53 (b) 65	(a) 71 (b) 82
(a) 17 (b) 28	0	0	0	0	0	q^3
(a) 34 (b) 46	0	0	0	q^2	0	$(q^{3/2}-q^{7/2})[2]^{-1/2}$
(a) 35 (b) 56	0	0	0	0	q^2	$(q^{3/2}-q^{7/2})\left(\frac{[3]}{[2]}\right)^{1/2}$
(a) 43 (b) 64	0	q^2	0	$1-q^4$	0	$(q^{11/2}-q^{7/2})[2]^{-1/2}$
(a) 53 (b) 65	0	0	q^2	0	0	$(q^{3/2}-q^{7/2})\left(\frac{[3]}{[2]}\right)^{1/2}$
(a) 71 (b) 82	q^3	$(q^{3/2}-q^{7/2})[2]^{-1/2}$	$(q^{3/2}-q^{7/2})\left(\frac{[3]}{[2]}\right)^{1/2}$	$(q^{11/2}-q^{7/2})[2]^{-1/2}$	$(q^{3/2}-q^{7/2})\left(\frac{[3]}{[2]}\right)^{1/2}$	q^4-2q^2+1

V. NEW LINK POLYNOMIAL

Since the QYBE without the spectral parameter is the same as the multiplication rule of the braid group, a monodromy representation of the braid group B_n can be obtained by the $\check{R}_q$ matrix,

$$D(b_i)=1\times 1\times\cdots\times 1\times\check{R}_q\times 1\times\cdots\times 1, \tag{8}$$

where the $\check{R}_q$ matrix is located in the ith and $(i+1)$th positions in the direct product.

From the monodromy representation a new link polynomial can be built up in terms of the standard way:[4,5,13]

$$\alpha(B,n)=(\tau\bar{\tau})^{-(n-1)/2}(\bar{\tau}/\tau)^{e(B)/2}\,\mathrm{Tr}(VD(B,n)) \tag{9}$$

where $B\in B_n$, V is a direct product of n matrices v,

$$V=v\times v\times\cdots\times v,$$

$$v=\mathrm{diag}(q^{-4},q^{-2},q^{-2},1,1,q^2,q^2,q^4)\times(q^{-4}+2q^{-2}+2+2q^2+q^4)^{-1}, \tag{10}$$

and

$$\tau=(1+2q^2+2q^4+2q^6+q^8)^{-1},\quad \bar{\tau}=q^8\tau. \tag{11}$$

From the eigenvalues of the $\check{R}_q$ matrix we have

$$(\check{R}_q-1)(\check{R}_q+q^2)(\check{R}_q^2-q^{10})(\check{R}_q-q^8)=0 \tag{12}$$

and the Alexander–Conway relation of the link polynomial

$$\begin{aligned}\alpha(Ab^5B,n)=&\,q^4(1-q^2+q^8)\alpha(Ab^4B,n)+q^{10}(1-q^6+2q^8)\alpha(Ab^3B,n)-q^{22}(2-q^2+q^8)\alpha(Ab^2B,n)\\&-q^{28}(1-q^6+q^8)\alpha(AbB,n)+q^{40}\alpha(AB,n)\\ \alpha(E,2)=&\,q^{-4}+2q^{-2}+2+2q^2+q^4.\end{aligned} \tag{13}$$

ACKNOWLEDGMENTS

The author would like to thank Professor Abdus Salam, the International Atomic Energy Agency, and UNESCO for the Hospitality at ICTP, Trieste, Italy.

This work was partly supported by the Natural Science Foundation of China through the Nankai Institute of Mathematics.

[1]C. N. Yang, Phys. Rev. Lett. **19**, 1312 (1967); R. J. Baxter, Ann. Phys. **70**, 193 (1972).

[2]L. Faddeev, *Integrable Models in 1+1-dimensional Quantum Field Theory*, in Les Houches Lecture 1982 (Elsevier, Amsterdam, 1984).

[3]L. Alvarez-Gaumé, C. Gomez, and G. Sierra, CERN-TH-5369/89; Phys. Lett. B 220, 142 (1989).

[4]V. F. R. Jones, Bull. Am. Math. Soc. **12**, 103 (1985).

[5]Y. Akutsu and M. Wadati, J. Phys. Soc. Jpn. **56**, 3039 (1987); Y. Akustu, T. Deguchi, and M. Wadati, J. Phys. Soc. Jpn. **56**, 3464 (1987); M. Wadati and Y. Akutsu, Prog. Theor. Phys. Supp. No. 94 (1988).

[6]R. P. Kulish, N. Yu. Reshetikhin, and E. K. Sklyanin, Lett. Math. Phys. **5**, 393 (1981).

[7]D. I. Olive and N. Turok, Nucl. Phys. B 220, 491 (1983).

[8]Bo-Yu Hou, Bo-Yuan Hou and Zhong-Qi Ma, Commun, Theor. Phys. **13**, 181 and 341 (1990).

[9]Z. Q. Ma, "Quantum sl(2) enveloping algebra and representations of braid group," in *Knots, Topology and Quantum Field Theories*, Proceedings of the Johns Hopkins Workshop on Current Problems in Particle Theory 13, Florence, Italy, edited by L. Lusanna (World Scientific, Singuapore, 1989), pp. 469–484.

[10]A. M. Kirillov and N. Yu. Reshetikhin, LOMI preprints, E-9-88, 1988.

[11]O. Babelon, H. J. de Vega and C. M. Viallet, Nucl. Phys. B **190**, 542 (1981); I. V. Cherednik, Theor. Math. Phys. **43**, 356 (1980); D. V. Chudnovsky and G. V. Chudnovsky, Phys. Lett A **79**, 36 (1980); C. L. Schultz, Phys. Rev. Lett. **46**, 629 (1981).

[12]M. Jimbo, Commun. Math. Phys. **102**, 537 (1986); Lett. Math. Phys. **11**, 247 (1986).

[13]Zhong-Qi Ma, J. Math. Phys. **31**, 550 (1990).

[14]J. J. de Swart, Rev. Mod. Phys. **35**, 916 (1963).

The Goursat problem for Maxwell's equations

Pierre Hillion
Institut Henri Poincaré, 75231 Paris, France

(Received 7 March 1990; accepted for publication 18 July 1990)

Using the spinor formalism of electromagnetism, the Goursat problem for Maxell's equations is defined and some examples of solutions are given.

I. INTRODUCTION

The Goursat problem[1] is a boundary value problem with data on the characteristics $\xi = z - x_0$, $\eta = z + x_0$ ($x_0 = ct$) of the 1-D wave equation $\partial_z^2\psi - \partial_{x_0}^2\psi = 0$. Some authors speak of a boundary value problem of the first kind[2,3] and of a characteristic initial value problem.[4]

In a previous work,[5] we pointed out how the Goursat problem may be generalized to the 3-D wave equation by considering the transverse variables x,y, as some parameters. Here, we consider the Goursat problem for Maxwell's equations.

Because the spinor formalism of electromagnetism is a powerful tool to discuss the Goursat problem, we start with a presentation of this formalism. We also define the electromagnetic modal solutions of Maxwell's equations.

II. SPINOR FORMALISM OF ELECTROMAGNETISM

Let us consider the propagation of an electromagnetic wave in the z direction of a homogeneous medium. To simplify, we assume $\varepsilon = \mu = 1$, where ε, μ are, respectively, the electric permittivity and the magnetic permeability. We now introduce the complex vector:

$$\Lambda = E + iH, \quad i = \sqrt{-1}, \tag{1}$$

where E, H are the electric and magnetic fields. In fact, Λ is a self-dual tensor and there exists[6] a well-known relationship between self-dual tensors and traceless spinors ψ_r^s of rank two ($r,s = 1,2$). One has

$$\begin{aligned} &\Lambda_x + i\Lambda_y = \psi_2^1, \\ &\Lambda_x - i\Lambda_y = \psi_1^2, \\ &\Lambda_z = \psi_1^1 = -\psi_2^2 \end{aligned} \tag{2}$$

Moreover if ψ_r^s is a solution of the Proca equation[7]

$$\begin{vmatrix} \partial_\eta & \partial_{\bar\xi} \\ \partial_\xi & -\partial_{\bar\xi} \end{vmatrix} \begin{vmatrix} \psi_1^1 & \psi_1^2 \\ \psi_2^1 & \psi_2^2 \end{vmatrix} = 0, \tag{3}$$

with

$$\begin{aligned} &\partial_\xi = \partial_z - \partial_{x_0}, \quad \partial_\eta = \partial_z + \partial_{x_0}, \\ &\partial_\zeta = \partial_x + i\partial_y, \quad \partial_{\bar\xi} = \partial_x - i\partial_y, \end{aligned} \tag{3'}$$

then it is easy to prove that Λ is a solution of the Maxwell equations.

Now let ψ be a solution of the scalar wave equation:

$$(\partial_\xi\partial_\eta + \partial_\zeta\partial_{\bar\xi})\psi = 0, \tag{4}$$

then the solution of the Proca equation (3) is

$$\begin{aligned} &\psi_2^1 = \partial_\zeta\partial_\eta\psi, \\ &\psi_1^2 = \partial_{\bar\zeta}\partial_\xi\psi, \\ &\psi_2^1 = \partial_\xi\partial_\eta\psi = -\partial_\zeta\partial_{\bar\zeta}\psi, \end{aligned} \tag{5}$$

which supplies, through the relations (2), the solutions of the Maxwell equations in terms of the scalar field ψ.

An interesting point to note is that Eq. (4) is just written in the form we previously used[5] to obtain the modal solutions of the scalar wave equation. They are

$$\psi^0 = \frac{1}{\tau_\mu} F\left(\frac{\omega u}{\tau_\mu}\right), \quad \psi^0_{i,j,k,l} = \partial_\xi^i\partial_\eta^j\partial_\xi^k\partial_{\bar\xi}^l\psi^0 \tag{6a}$$

and

$$\psi^1 = \frac{1}{\tau_\mu} F\left(\frac{\tau_\mu}{\omega u}\right), \quad \psi^1_{i,j,k,l} = \partial_\xi^i\partial_\eta^j\partial_\xi^k\partial_{\bar\xi}^l\psi^1. \tag{6b}$$

In these expressions, one has

$$\begin{aligned} &u = (a+\xi)(b+\eta) + (c+\zeta)(d+\bar\zeta), \\ &\tau_1 = a+\xi, \ \tau_2 = b+\eta, \ \tau_3 = c+\zeta, \ \tau_4 = d+\bar\zeta, \end{aligned} \tag{7}$$

a,b,c,d,ω are some constants, F is an arbitrary function with sufficient derivatives, ∂ denotes a derivative and i,j,k,l are non-negative numbers. The fields ψ^0, ψ^1 are the fundamental modes, $\psi^0_{i,j,k,l}$ and $\psi^1_{i,j,k,l}$ the higher-order modes.

The relations (5) and (6) supply the modal solutions of the Maxwell equations:

$$\begin{aligned} &\psi_2^{\alpha,1} = \psi^\alpha_{i,j+1,k+1,l}, \\ &\psi_1^{\alpha,2} = \psi^\alpha_{i+1,j,k,l+1}, \quad \alpha = 0,1. \\ &\psi_1^{\alpha,1} = \psi^\alpha_{i+1,j+1,k,l}, \end{aligned} \tag{8}$$

The best known of these modal solutions are the focus wave modes[8,9] obtained with $\alpha = 0$ when F is an exponential function and $\tau_\mu = \tau_1$ (or τ_2).

Let us give another example. Starting with $\psi^0 = u^{-1}$. We get from (8) for the fundamental mode:

$$\begin{aligned} &\psi_2^1 = (2/u^3)(a+\xi)(d+\bar\zeta), \\ &\psi_2^1 = (2/u^3)(b+\eta)(c+\zeta), \\ &\psi_1^1 = (1/u^3)\{(b+\eta)(a+\xi) - (c+\zeta)(d+\bar\zeta)\}. \end{aligned} \tag{9}$$

It appears that in a homogeneous medium the Maxwell equations are very rich in modal solutions.

III. THE GOURSAT PROBLEM

For the homogeneous wave equation (4) the Goursat problem is defined[5] by the boundary data:

 0022-2488/90/123085-04$03.00

$$\psi(\zeta,\bar{\zeta},0,\eta)=\varphi_1(\zeta\bar{\zeta},\eta),\quad \psi(\zeta,\bar{\zeta},\xi,0)=\varphi_2(\zeta,\bar{\zeta},\xi), \tag{10}$$

with the compatibility condition,

$$\psi(\zeta,\bar{\zeta},0,\eta)=\varphi_1(\zeta\bar{\zeta},\eta),\quad \psi(\zeta,\bar{\zeta},\xi,0)=\varphi_2(\zeta,\bar{\zeta},\xi) \tag{10'}$$

Similarly, we define the Goursat problem for the Proca equation (3) by the boundary data:

$$\psi(\zeta,\bar{\zeta},0,\eta)=\varphi_1(\zeta\bar{\zeta},\eta),\quad \psi(\zeta,\bar{\zeta},\xi,0)=\varphi_2(\zeta,\bar{\zeta},\xi), \tag{11}$$

with the compatibility conditions

$$\varphi^s_r(\zeta,\bar{\zeta},0)=\vartheta^s_r(\zeta,\bar{\zeta},0),\quad r,s=1,2. \tag{11'}$$

But here the boundary data must satisfy supplementary constraints easy to find from (3):

$$\partial_\eta\varphi^1_1+\partial_{\bar{\zeta}}\varphi^1_2=0,\quad \partial_{\bar{\zeta}}\varphi^1_1-\partial_\eta\varphi^2_1=0, \tag{12a}$$

$$\partial_\xi\vartheta^1_1+\partial_\zeta\vartheta^2_1=0,\quad \partial_\zeta\vartheta^1_1-\partial_\xi\vartheta^1_2=0, \tag{12b}$$

leading to

$$\varphi^1_1=\partial_\eta\partial_{\bar{\zeta}}\varphi,\quad \varphi^1_2=-\partial^2_\eta\varphi,\quad \varphi^2_1=\partial^2_{\bar{\zeta}}\varphi, \tag{13a}$$

$$\vartheta^1_1=\partial_\xi\partial_\zeta\vartheta,\quad \vartheta^2_1=-\partial^2_\xi\vartheta,\quad \vartheta^1_2=\partial^2_\zeta\vartheta, \tag{13b}$$

where φ is a function of ζ, $\bar{\zeta}$, η, and ϑ a function of ζ, $\bar{\zeta}$, ξ.

To sum up, the Goursat problem for the Proca equation is characterized by the boundary data (11), (11'), and (13).

As an important result, the modal waves (8) satisfy these boundary conditions. Consequently, these waves are solutions of some Goursat problem. Let us prove this result for the fundamental mode ψ^0 with $\tau_\mu=\tau_1$. One has

$$\psi^0=\frac{1}{a+\xi}F(v),\quad v=\omega\left(b+\eta\frac{(c+\zeta)(d+\bar{\zeta})}{a+\xi}\right), \tag{14}$$

then F', F'' being the first and second derivatives of F with respect to v, we get from (8) and (14)

$$\begin{aligned}
\psi^1_2&=\omega^2\frac{(d+\bar{\zeta})}{(a+\xi)^2}F'',\\
\psi^2_1&=-2\omega\frac{(c+\zeta)}{(a+\xi)^3}F'-\omega^2\frac{(c+\zeta)^2(d+\bar{\zeta})}{(a+\xi)^4}F'',\\
\psi^1_1&=\frac{-\omega F'}{(a+\xi)^2}-\omega^2\frac{(c+\zeta)(d+\bar{\zeta})}{(a+\xi)^3}F'',
\end{aligned} \tag{15}$$

and using the notations:

$$\begin{aligned}
\tilde{v}&=\omega\left(b+\eta+\frac{(c+\zeta)(d+\bar{\zeta})}{a}\right),\\
\hat{v}&=\omega\left(b+\frac{(c+\zeta)(d+\bar{\zeta})}{a+\xi}\right),
\end{aligned} \tag{16}$$

we deduce from (15) on $\eta=0$:

$$\begin{aligned}
\vartheta^1_2&=\omega^2\frac{(d+\bar{\zeta})}{(a+\xi)^2}F''(\hat{v}),\\
\vartheta^2_1&=-2\omega\frac{(c+\zeta)}{(a+\xi)^3}F'(\hat{v})\\
&\quad-\omega^2\frac{(c+\zeta)^2(d+\bar{\zeta})}{(a+\xi)^4}F''(\hat{v}),\\
\vartheta^1_1&=-\omega\frac{F'(\hat{v})}{(a+\xi)^2}-\frac{\omega^2}{(a+\xi)^3}(c+\zeta)(d+\bar{\zeta})F''(\hat{v}),
\end{aligned} \tag{17a}$$

and on $\xi=0$

$$\begin{aligned}
\varphi^1_2&=\frac{\omega^2}{a^2}(d+\bar{\zeta})F''(\tilde{v}),\\
\varphi^2_1&=-\frac{2\omega}{a^3}(c+\zeta)F'(\tilde{v})\\
&\quad-\frac{\omega^2}{a^4}(c+\zeta)^2(d+\bar{\zeta})F''(\tilde{v}),\\
\varphi^1_1&=-\frac{\omega}{a^2}F'(\tilde{v})-\frac{\omega^2}{a^3}(c+\zeta)(d+\bar{\zeta})F''(\tilde{v}).
\end{aligned} \tag{17b}$$

One sees at once that the compatibility conditions (11') are satisfied as well as the relations (13) with

$$\varphi=[(d+\bar{\zeta})/a^2]F(\tilde{v}),\quad \vartheta=[1/(d+\bar{\zeta})]F(\hat{v}). \tag{18}$$

Let us remark that in (17) we used the equalities:

$$\partial_{\tilde{v}}F(\tilde{v})=\{\partial_v F(v)\}_{v=\tilde{v}},\quad \partial_{\hat{v}}F(\tilde{v})=\{\partial_v F(v)\}_{v=\hat{v}} \tag{19}$$

and similar relations for the second derivatives.

We discuss in the next section another class of Goursat problems where solutions are easy to find.

IV. THE SEPARABLE GOURSAT PROBLEM

The separable Goursat problem is defined by the boundary data:

$$\varphi^s_r=f^s_r(\zeta,\bar{\zeta})g^s_r(\eta),\quad \vartheta^s_r=f^s_r(\zeta,\bar{\zeta})h^s_r(\xi),\quad r,s=1,2, \tag{20}$$

with the compatibility conditions:

$$g^s_r(0)=h^s_r(0),\quad r,s=1,2, \tag{20'}$$

while the relations (12) take the form:

$$\begin{aligned}
&\lambda g^1_2=\partial_\eta g^1_1,\quad \mu g^1_1=\partial_\eta g^2_1\\
&\rho h^2_1=\partial_\xi h^1_1,\quad \sigma h^1_1=\partial_\xi h^1_2
\end{aligned} \tag{21a}$$

$$\begin{aligned}
&\lambda f^1_1=\partial_{\bar{\zeta}}f^1_2,\quad \mu f^2_1=\partial_{\bar{\zeta}}f^1_1\\
&\rho f^1_1=\partial_\zeta f^1_1,\quad \sigma f^1_2=\partial_\zeta f^1_1
\end{aligned} \tag{21b}$$

where λ, μ, ζ, σ are some constants. From (21b) we deduce $\lambda\sigma=\rho\mu$ while f^s_r satisfies the equation:

$$(\partial_\zeta\partial_{\bar{\zeta}}+\lambda\sigma)f^s_r=0,\quad r,s=1,2, \tag{22}$$

which was the elementary solutions:

$$f=e^{i(\lambda\bar{\zeta}+\sigma\zeta)},\quad f=J_0(\sqrt{\lambda\sigma\zeta\bar{\zeta}}), \tag{22'}$$

where J_0 is the Bessel function of the first kind of order zero.

Taking (21a) into account, the conditions (20') become:

$$\begin{aligned}
&\rho\sigma g^2_1(0)-\partial_\xi h^1_2(0)=0,\quad \sigma\partial_\eta g^2_1(0)-\mu\partial h^1_2(0)=0,\\
&\mu\lambda h^1_2(0)-\partial^2_\eta g^2_1(0)=0.
\end{aligned} \tag{23}$$

We now discuss some particular separable Goursat problems starting with the case where $g^1_1(\eta)=h^1_1(\xi)=1$. According to (21) this implies:

$$\begin{aligned}
&g^1_1=h^2_1=0,\quad g^2_1=\mu\eta,\quad h^1_2=\sigma\xi,\\
&f^2_1=(1/\mu)\partial_{\bar{\zeta}}f^1_1,\quad f^1_2=(1/\sigma)\partial_\zeta f^1_1,
\end{aligned} \tag{24}$$

so that the boundary conditions become

$$\varphi_2^1 = 0, \qquad \vartheta_2^1 = \xi \partial_\zeta f_1^1,$$
$$\varphi_1^2 = \eta \partial_{\bar\zeta} f_1^1, \quad \vartheta_1^2 = 0, \tag{25}$$
$$\varphi_1^1 = f_1^1, \qquad \vartheta_1^1 = f_1^1,$$

where f_1^1 is a solution of Eq. (22).

To solve this Goursat problem one has just to remark that ψ_1^1 satisfies the wave equation (4). Consequently, it is enough to look for the solution of Eq. (4) with the boundary conditions:

$$\psi_1^1(\zeta,\bar\zeta,0,\eta) = \psi_1^1(\zeta,\bar\zeta,\xi,0) = f_1^1(\zeta,\bar\zeta). \tag{26}$$

The obvious solutions for the elementary functions (22') are

$$\psi_1^1 = I_0(k\sqrt{\xi\eta}) J_0(k\sqrt{\zeta\bar\zeta}), \qquad k^2 = \lambda\sigma, \tag{27a}$$
$$\psi_1^1 = I_0(k\sqrt{\xi\eta}) e^{i(\lambda\bar\zeta + \sigma\zeta)}, \tag{27b}$$

where I_0 is the modified Bessel function of the first kind of order zero.

Now according to (5) one has $\psi_1^1 = \partial_\xi \partial_\eta \psi$ so that we get from (27) and from the well-known properties of the Bessel functions $\psi = 4k^{-2}\psi_1^1$. Substituting this last result into (5) gives the solutions in terms of the Bessel functions:

$$\psi_2^1 = -\left(\frac{\xi\bar\xi}{\eta\zeta}\right)^{1/2} I_1(k\sqrt{\xi\eta}) J_1(k\sqrt{\zeta\bar\zeta}),$$
$$\psi_1^2 = -\left(\frac{\eta\zeta}{\xi\bar\xi}\right)^{1/2} I_1(k\sqrt{\xi\eta}) J_1(k\sqrt{\zeta\bar\zeta}), \tag{28a}$$
$$\psi_1^1 = I_0(k\sqrt{\xi\eta}) J_0(k\sqrt{\zeta\bar\zeta}),$$
$$\psi_2^1 = 2i\left(\frac{\sigma\xi}{\lambda\eta}\right) I_1(k\sqrt{\xi\eta}) e^{i(\lambda\bar\zeta + \sigma\zeta)},$$
$$\psi_1^2 = 2i\left(\frac{\lambda\eta}{\sigma\xi}\right) I_1(k\sqrt{\xi\eta}) e^{i(\lambda\bar\zeta + \sigma\zeta)}, \tag{28b}$$
$$\psi_1^1 = I_0(k\sqrt{\xi\eta}) e^{i(\lambda\bar\zeta + \sigma\zeta)};.$$

The expressions (28) are the solutions of the Goursat problem (25) for the Proca equation (3) when f_1^1 is an elementary function (22').

Let us now assume that f_1^1 is a function with the Fourier expansion:

$$f_1^1(\zeta,\bar\zeta) = \sum_{\lambda,\sigma} c_{\lambda\sigma} e^{i(\lambda\bar\zeta + \sigma\zeta)} \tag{29a}$$

or with the Fourier–Bessel expansion:[10]

$$f_1^1(r) = \sum_{m=1}^{\infty} a_m J_0(kj_m r), \quad r^2 = \zeta\bar\zeta, \tag{29b}$$

where $j_0, j_1, j_2 \ldots$ are the positive zeros of J_0 arranged in ascending order of amplitude. Then the solutions ψ_1^1 of Eq. (4) become for these boundary conditions f_1^1:

$$\psi_1^1 = \sum_{\lambda\sigma} c_{\lambda\sigma} I_0(k\sqrt{\xi\eta}) e^{i(\lambda\bar\zeta + \sigma\zeta)}, \tag{30a}$$

$$\psi_1^1 = \sum_{m=1}^{\infty} a_m I_0(kj_m\sqrt{\xi\eta}) J_0(kj_m r). \tag{30b}$$

As was obtained previously from the third relation (5) we get

$$\psi = 4\sum_{\lambda,\sigma} \frac{c_{\lambda\sigma}}{\lambda\sigma} I_0(k\sqrt{\xi\eta}) e^{i(\lambda\bar\zeta + \sigma\zeta)}, \tag{31a}$$

$$\psi = \frac{4}{k^2} \sum_{m=1}^{\infty} \frac{a_m}{j_m^2} I_0(kj_m\sqrt{\xi\eta}) J_0(kj_m r), \tag{31b}$$

Substituting (31) into (5) supplies the solution of the Goursat problem provided that all the series converge:

$$\psi_2^1 = 2i\sqrt{\frac{\xi}{\eta}} \sum_{\lambda\sigma} c_{\lambda\sigma} \sqrt{\frac{\sigma}{\lambda}} I_1(k\sqrt{\xi\eta}) e^{i(\lambda\bar\zeta + \sigma\zeta)},$$
$$\psi_1^2 = 2i\sqrt{\frac{\eta}{\xi}} \sum_{\lambda\sigma} c_{\lambda\sigma} \sqrt{\frac{\lambda}{\sigma}} I_1(k\sqrt{\xi\eta}) e^{i(\lambda\bar\zeta + \sigma\zeta)}, \tag{32a}$$
$$\psi_1^1 = \sum_{\lambda\sigma} c_{\lambda\sigma} I_0(k\sqrt{\xi\eta}) e^{i(\lambda\bar\zeta + \sigma\zeta)},$$

and

$$\psi_2^1 = -\sqrt{\frac{\xi\bar\xi}{\eta\zeta}} \sum_{m=1}^{\infty} a_m I_1(kj_m\sqrt{\xi\eta}) J_1(kj_m r),$$
$$\psi_1^2 = -\sqrt{\frac{\eta\zeta}{\xi\bar\xi}} \sum_{m=1}^{\infty} a_m I_1(kj_m\sqrt{\xi\eta}) J_1(kj_m r), \tag{32b}$$
$$\psi_1^1 = \sum_{m=1}^{\infty} a_m I_0(kj_m\sqrt{\xi\eta}) J_0(kj_m r).$$

As a second example of separable Goursat problem, let us consider the case $\lambda = \mu$ which implies $\rho = \sigma$ and from (21a) and (23):

$$g_1^1 = g_2^1 = g_1^2 = e^{\lambda\eta}, \quad h_1^1 = h_2^1 = h_1^2 = e^{\sigma\xi}, \tag{33}$$

leading to the boundary conditions:

$$\varphi_2^1 = (1/\lambda) e^{\lambda\eta} \partial_\zeta f_1^1, \quad \vartheta_2^1 = (1/\sigma) e^{\sigma\xi} \partial_\zeta f_1^1,$$
$$\varphi_1^2 = (1/\lambda) e^{\lambda\eta} \partial_{\bar\zeta} f_1^1, \quad \vartheta_1^2 = (1/\sigma) e^{\sigma\xi} \partial_{\bar\zeta} f_1^1, \tag{34}$$
$$\varphi_1^1 = e^{\lambda\eta} f_1^1, \qquad \vartheta_1^1 = e^{\sigma\xi} f_1^1,$$

where f_1^1 is still a solution of Eq. (22). For the elementary solution $f_1^1 = e^{i(\lambda\bar\zeta + \sigma\zeta)}$ the boundary conditions (34) reduce to

$$\varphi_1^1 = i\varphi_2^1 = -i\varphi_2^1 = e^{\lambda\eta} e^{i(\lambda\bar\zeta + \sigma\zeta)},$$
$$\vartheta_1^1 = -i\vartheta_2^1 = -i\vartheta_1^2 = e^{\sigma\xi} e^{i(\lambda\bar\zeta + \sigma\zeta)}, \tag{35}$$

and the solution of the Goursat problem is obtained at once:

$$\psi_1^1 = -i\psi_2^1 = -i\psi_1^2 = e^{(\lambda\eta + \sigma\xi)} e^{i(\lambda\bar\zeta + \sigma\zeta)}. \tag{36}$$

For the other elementary solutions (22'), the boundary data become

$$\varphi_2^1 = -\frac{1}{2} e^{\lambda\eta} \sqrt{\frac{\lambda\bar\zeta}{\sigma\zeta}} J_1(\sqrt{\lambda\sigma\zeta\bar\zeta}), \quad \vartheta_2^1 = -\frac{1}{2} e^{\sigma\xi} \sqrt{\frac{\lambda\bar\zeta}{\sigma\zeta}} J_1(\lambda\sigma\zeta\bar\zeta),$$
$$\varphi_1^2 = -\frac{1}{2} e^{\lambda\eta} \sqrt{\frac{\sigma\zeta}{\lambda\bar\zeta}} J_1(\sqrt{\lambda\sigma\zeta\bar\zeta}), \quad \vartheta_1^2 = -\frac{1}{2} e^{\sigma\xi} \sqrt{\frac{\sigma\zeta}{\lambda\bar\zeta}} J_1(\sqrt{\lambda\sigma\zeta\bar\zeta}), \tag{37}$$
$$\varphi_1^1 = e^{\lambda\eta} J_0(\sqrt{\lambda\sigma\zeta\bar\zeta}), \qquad \vartheta_1^1 = e^{\sigma\xi} J_0(\sqrt{\lambda\sigma\zeta\bar\zeta}).$$

Proceeding as we did previously, we find as a solution of the Goursat problem

$$\psi_2^1 = -\frac{1}{2}\sqrt{\frac{\lambda\bar{\xi}}{\sigma\zeta}}\, e^{(\lambda\eta+\sigma\xi)} J_1(\sqrt{\lambda\sigma\xi\bar{\xi}}),$$
$$\psi_1^2 = -\frac{1}{2}\sqrt{\frac{\sigma\zeta}{\lambda\bar{\xi}}}\, e^{(\lambda\eta+\sigma\xi)} J_1(\sqrt{\lambda\sigma\xi\bar{\xi}}), \qquad (38)$$
$$\psi_1^1 = e^{(\lambda\eta+\sigma\xi)} J_0(\sqrt{\lambda\sigma\xi\bar{\xi}}).$$

One notes at once the similarity of (38) and (28b) which comes from the invariance of Eq. (4) under the transformations $(\xi,\eta)\leftrightarrow(\zeta,\bar{\zeta})$ and $\xi\leftrightarrow\eta$, $\zeta\leftrightarrow\bar{\zeta}$.

As shown previously, if f_1^1 is given by (29a) or (29b), one may use similar expansions to (32a) and (32b) with the expressions (36) and (38) taking the place of (28a) and (28b).

As a final example of a separable Goursat problem, let us remark that relations (13) suggest to consider boundary conditions with functions φ, ϑ, such as

$$\varphi = f(\lambda\bar{\xi})g(\lambda\eta-\zeta), \quad \vartheta = -f(\xi+\lambda\bar{\zeta})g(-\zeta). \qquad (39)$$

Denoting f', f'' and g', g'' the first and second derivatives of f, g with respect to the variables $\xi+\lambda\bar{\xi}$ and $\lambda\eta-\zeta$, respectively, relations (13) become

$$\varphi_1^1 = \lambda^2 f'(\lambda\bar{\xi})g'(\lambda\eta-\zeta),$$
$$\varphi_1^2 = \lambda^2 f''(\lambda\bar{\xi})g(\lambda\eta-\zeta), \qquad (40a)$$
$$\varphi_2^1 = -\lambda^2 f(\lambda\bar{\xi})g''(\lambda\eta-\zeta),$$
$$\vartheta_1^1 = \lambda^2 f'(\xi+\lambda\bar{\xi})g'(-\zeta),$$
$$\vartheta_1^2 = \lambda^2 f''(\xi+\lambda\bar{\xi})g(-\zeta), \qquad (40b)$$
$$\vartheta_2^1 = -\lambda^2 f(\xi+\lambda\bar{\xi})g''(-\zeta).$$

The compatability conditions (11′) are fulfilled and using the same technique as previously we get as a solution of this Goursat problem:

$$\psi_2^1 = -\lambda^2 f(\xi+\lambda\bar{\xi})g''(\lambda\eta-\zeta),$$
$$\psi_1^2 = \lambda^2 f''(\xi+\lambda\bar{\xi})g(\lambda\eta-\zeta), \qquad (41)$$
$$\psi_1^1 = \lambda^2 f'(\xi+\lambda\bar{\xi})g'(\lambda\eta-\zeta).$$

With (41), one may obtain easily the solution of the Goursat problem when φ and ϑ have the expansions:

$$\varphi = \sum_\lambda c_\lambda f(\lambda\bar{\xi})g(\lambda\eta-\zeta), \quad \vartheta = \sum_\lambda c_\lambda f(\xi+\lambda\bar{\xi})g(-\zeta). \qquad (42)$$

V. CONCLUSIONS

This work proves that the spinor formalism of electromagnetism together with the Proca equation is a powerful tool to solve the Goursat problem for Maxwell's equations. Here, we only defined the Goursat problem giving some examples of solutions in the class of modal waves or for separable problems. But we left untouched some important mathematical questions such as conditions for the existence and uniqueness of solutions as well as the quest for (numerical?) methods to solve this problem. Clearly, further work is needed.

One also has from a physical point of view to discuss the meaning of the Goursat problem that could appear as some relativistic generalization of the Huyghens principle stating that the sources may be replaced by boundary conditions on the characteristics. Such a situation emerges naturally from the relativistic front form of electromagnetism initiated by Dirac[11] many years ago. Hopefully, one expects that some solutions discussed in this work will be useful for the later developments of this theory.

[1] A. N. Tychonov and A. A. Samarski, *Partial Differential Equations of Mathematical Physics* (Holden-Day, San Francisco, CA, 1964).
[2] S. L. Sobolev, *Partial Differential Equations of Mathematical Physics* (Pergamon, New York, 1964).
[3] I. N. Sneddon, *Elements of Partial Differential Equations* (MacGraw-Hill, New York, 1957).
[4] R. Courant and D. Hilbert, *Methods of Mathematical Physics—Vol. II* (Interscience, New York, 1962).
[5] P. Hillion, J. Math. Phys. **31**, 1939 (1990).
[6] E. Cartan, *Leçons sur la Théorie des Spineurs* (Hermann, Paris, 1938).
[7] E. M. Corson, *Introduction to Tensors, Spinors and Relativistic Wave Equations* (Blackic, London, 1953).
[8] J. N. Brittingham, J. Appl. Phys. **54**, 1179 (1983).
[9] P. Hillion, J. Appl. Phys. **60**, 981 (1986).
[10] G. N. Watson, *Theory of Bessel Functions* (Cambridge, U.P., Cambridge, 1962).
[11] P. A. M. Dirac, Rev. Mod. Phys. **21**, 392 (1949).