Canonical transforms, separation of variables, and similarity solutions for a class of parabolic differential equations

Kurt Bernardo Wolf

Centro de Investigación en Matemáticas Aplicadas y en Sistemas (CIMAS), Universidad Nacional Autónoma de México, México D.F., Mexico
(Received 20 June 1975)

Using the method of canonical transforms, we explicitly find the similarity or kinematical symmetry group, all "separating" coordinates and invariant boundaries for a class of differential equations of the form $[\alpha\,\partial^2/\partial q^2+\beta q\,\partial/\partial q+\gamma q^2+\delta q+\epsilon\,\partial/\partial q+\zeta\,]\,u(q,t)=-i(\partial/\partial t)\,u(q,t)$, or of the form $[\alpha'(\partial^2/\partial q^2+\mu/q^2)+\beta' q\,\partial/\partial q+\gamma' q^2\,]\,u(q,t)=-i(\partial/\partial t)\,u(q,t)$, for complex $\alpha,\beta,\ldots,\gamma'$. The first case allows a six-parameter WSL(2,R) invariance group and the second allows a four-parameter O(2)$\otimes$ SL(2,R) group. Any such differential equation has an invariant scalar product form which, in the case of the heat equation, appears to be new. The proposed method allows us to work with the group, rather than the algebra, and reduces all computation to the use of 2×2 matrices.

I. INTRODUCTION

A. In a recent series of papers[1-3] we have dealt with realizations of Lie algebras in terms of second-order differential operators and their exponentiation to the group. In contradistinction with first-order differential realizations, which produce *geometric* transformations of the general form

$$f(q)\xrightarrow{h} f_h(q)=\mu(q,h)f[\bar{q}_h(q)], \tag{1.1}$$

where μ is a multiplier function, second-order differential operators, when exponentiated, will in general lead to an *integral transform*

$$f(q)\xrightarrow{g} f_g(q)=\int dq'\,K_g(q,q')\,f(q'), \tag{1.2}$$

where $K_g(q,q')$ is an integral kernel. The action (1.1) has been extensively treated[4] since the times of Lie,[5] while only recently[6,7] have forms (1.2) been subjected to intensive study. In Refs. 1 and 2, we have worked with the groups SL(2,C) [the group of unimodular 2×2 complex matrices] and the associated mappings (1.2) as unitary transformations between Hilbert spaces, one of them being $\mathcal{L}^2(R)$ or $\mathcal{L}^2(R^+)$ (Lebesgue square-integrable functions on the real line R or on the positive half-line R^+), and the other one, a space of analytic functions over regions of the complex plane *à la* Bargmann.[8] When the mapping (1.2) belongs to the SL(2,R) subgroup (of unimodular 2×2 real matrices), the "Bargmann" spaces collapse to ordinary $\mathcal{L}^2$ spaces. We have called these mappings *canonical transforms* since they arose from the study of complex canonical transformations in quantum mechanics. They include as particular cases the transforms of Fourier, Laplace, Weierstrass, Bargmann, Hankel, and Barut–Girardello.

B. If H is a second-order differential operator in a variable q, element of a Lie algebra (which in this paper will be sl(2,R) or wsl(2,R)—semidirect sum of the Weyl and sl(2,R) algebras), the solution of the parabolic differential equation

$$\theta H u(q,t)=-i\,\frac{\partial}{\partial t}U(q,t), \tag{1.3}$$

where θ is an in general complex constant, can be expressed as a *canonical transform* of the initial condition

$$u(q)\equiv u(q,0),$$

$$u(q,t)=\exp(it\theta H)\,u(q). \tag{1.4}$$

Now we can subject $u(q)$ to a general integral transform (1.2) to a $u_g(q)$, and the corresponding $u_g(q,t)$ will still be a solution of (1.3) and, in fact, a geometric transform of $u(q,t)$. This will be the group of symmetries, kinematical[9,10] or similarity[11,12] group of the differential equation (1.3). We can further look for the invariant lines (boundaries) under g_0 in the q–t plane, $v(q,t)$ and thus use the generator of the said transformation to separate Eq. (1.3) into two ordinary differential equations, one in v and one in t. The solution of (1.3) will then have the form of a general superposition of *separable* solutions,[13-17]

$$u_s(v(q,t),t)=\exp[iS(v,t)]\,V_s(v)\,T_s(t), \tag{1.5}$$

where $S(v,t)$ is a *multiplier* function (*not* expressible as a function in v plus a function in t).

C. Our claim in this article is that we can considerably simplify the process of finding these features for the class of differential equations (1.3) by starting with a given group and pair its realization in terms of a Lie algebra of second-order differential operators with the *matrix realization* of the group. Since we shall be dealing with real subgroups of WSL(2,C), the algebra required is essentially that of 2×2 matrices. This can be used to replace the rather lengthy conventional methods for finding separable coordinates and similarity groups through the solution of partial and coupled differential equations and the exhaustive examination of multiparameter ranges.

The canonical transform method, as used here, has the following limitations: it applies only to differential equations where H in (1.3) is of the form

$$H=\alpha\,\frac{d^2}{dq^2}+\beta q\,\frac{d}{dq}+\gamma q^2+\delta q+\epsilon\,\frac{d}{dq}+\zeta, \tag{1.6a}$$

or of the form

$$H=\alpha'\left(\frac{d^2}{dq^2}+\frac{\mu}{q^2}\right)+\beta' q\,\frac{d}{dq}+\gamma' q^2, \tag{1.6b}$$

for complex $\alpha,\beta,\ldots,\gamma'$, i.e., it applies only to a

particular class of parabolic, linear, second-order differential equations. Yet this class contains the physically interesting cases of the heat equation and the Schrödinger equations for the free particle or quadratic (attractive or repulsive) plus linear or inverse-quadratic potentials in one dimension. Through a simple point transformation, these can be related to the pseudo-Coulomb Schrödinger equation.[3] Our tabulated results are exhaustive within the group framework.[18]

D. The outline of the paper is the following. In Sec. II we assemble the mathematical tools: the algebra and group realizations in terms of second-order differential operators (1.6a) and their exponentiation to the six-parameter group, as acting on the space $\mathcal{L}^2(R)$ of functions and its adjoint action on the algebra; eigenfunctions and their eigenvalues for any operator in the algebra can thus be found in terms of their *orbit representatives*. In Sec. III we allow for the complexification of the group, and phrase the solution of (1.3) in terms of canonical transforms, reducing the problem of finding separating coordinates associated with a second operator in the algebra, to the manipulation of 2×2 matrices. We exemplify some of these developments for the heat equation as a complex canonical transform, pointing out the existence of a new quadratic—scalar product—invariant. Some of the group-integrated features of similarity methods are seen in Sec. IV. The free particle and heat equation are used as examples. In the latter, the set of bounded transformations constitute a semigroup. In Sec. V, differential equations with operators of the class (1.6b) are treated. Some connections, conclusions, and directions for further work are collected in Sec. VI.

II. THE GROUP WSL (2,R) AND ITS ORBIT STRUCTURE

A. The Heisenberg–Weyl algebra[19] w, of generators Q, P, and $\mathbb{1}$ is defined through the commutator brackets

$$[Q,P]=i\mathbb{1}, \quad [Q,\mathbb{1}]=0, \quad [P,\mathbb{1}]=0. \tag{2.1}$$

On the Hilbert space $\mathcal{L}^2(R)$, it is known[20] that every representation of w is unitarily equivalent to the Schrödinger representation

$$Qf(q)=qf(q), \quad Pf(q)=-i\frac{d}{dq}f(q), \quad \mathbb{1}f(q)=f(q), \tag{2.2}$$

which is densely defined and self-adjoint in $\mathcal{L}^2(R)$. The generator $\mathbb{1}$ is in the center of the algebra and thus denoted as the identity operator to start with.

B. We can exponentiate (2.2) to a unitary representation of the Weyl group w, where the elements[19] $\omega(x,y,z)\in W$ act on $f\in\mathcal{L}^2(R)$ as,

$$[T_\omega(x,y,z)f](q)\equiv\{\exp[i(xQ+yP+z\mathbb{1})f]\}(q) = \exp[i(xq+\tfrac{1}{2}xy+z)]f(q+y). \tag{2.3}$$

Defining for convenience $\xi=(x,y)$ as a two-component row vector, its transpose $\xi^T=\binom{x}{y}$ and $\Omega=\left(\begin{smallmatrix}0&-1\\1&0\end{smallmatrix}\right)$, the product law in W can be written, for $\omega(\xi,z)\equiv\omega(x,y,z)$ as,

$$\omega(\xi_1,z_1)\omega(\xi_2,z_2)=\omega(\xi_1+\xi_2,\, z_1+z_2+\tfrac{1}{2}\xi_1\Omega\xi_2^T), \tag{2.4}$$

so that the group indentity is $\omega(0,0)$ and $\omega(\xi,z)^{-1}=\omega(-\xi,-z)$.

C. Out of the enveloping algebra $\bar{w}$ of w, we want to produce other Lie algebras under the commutator bracket. The set of second-order expressions,

$$I_1=\tfrac{1}{4}(P^2-Q^2), \quad I_2=\tfrac{1}{4}(QP+PQ), \quad I_3=\tfrac{1}{4}(P^2+Q^2), \tag{2.5}$$

are densely defined and self-adjoint on $\mathcal{L}^2(R)$, satisfying,

$$[I_1,I_2]=-iI_3, \quad [I_3,I_1]=iI_2, \quad [I_2,I_3]=iI_1, \tag{2.6}$$

which we recognize as the $\mathrm{sl}(2,R)\cong\mathrm{su}(1,1)\cong\mathrm{so}(2,1)\cong\mathrm{sp}(2,R)$ algebra.[4] No other unitarily inequivalent, finite-dimensional algebra of finite-order expressions can be found in $\bar{w}$ besides (2.1), (2.5), and their composition.[21]

D. The algebra (2.5) can be exponentiated to the group $\mathrm{SL}(2,R)$ of real unimodular 2×2 matrices through its one-parameter subgroups,

$$\exp(i\alpha I_1): \begin{pmatrix}\cosh\frac{1}{2}\alpha & -\sinh\frac{1}{2}\alpha\\ -\sinh\frac{1}{2}\alpha & \cosh\frac{1}{2}\alpha\end{pmatrix}, \tag{27a}$$

$$\exp(i\beta I_2): \begin{pmatrix}\exp(-\frac{1}{2}\beta) & 0\\ 0 & \exp(\frac{1}{2}\beta)\end{pmatrix}, \tag{2.7b}$$

$$\exp(i\gamma I_3): \begin{pmatrix}\cos\frac{1}{2}\gamma & -\sin\frac{1}{2}\gamma\\ \sin\frac{1}{2}\gamma & \cos\frac{1}{2}\gamma\end{pmatrix}, \tag{2.7c}$$

$$\exp(ic\tfrac{1}{2}Q^2): \begin{pmatrix}1 & 0\\ c & 1\end{pmatrix}, \tag{2.7d}$$

$$\exp(ib\tfrac{1}{2}P^2): \begin{pmatrix}1 & -b\\ 0 & 1\end{pmatrix}, \tag{2.7e}$$

so that every $\mathbf{A}\equiv\left(\begin{smallmatrix}a&b\\c&d\end{smallmatrix}\right)\in\mathrm{SL}(2,R)$ [with $ad-bc=1$ for unimodularity] can be decomposed in terms of two or more of the elements (2.7). Now, the representation of $\mathrm{sl}(2,R)$ on $\mathcal{L}^2(R)$ obtained from (2.2) can also be exponentiated to a unitary representation of $\mathrm{SL}(2,R)$ on the same space as[1,6,14]

$$[C\left(\begin{smallmatrix}a&b\\c&d\end{smallmatrix}\right)f](q)=\int_R dq'A(q,q')f(q') = (2\pi b)^{-1/2}\exp(-i\pi/4)\int_{-\infty}^{\infty}dq' \times\exp[(i/2b)(aq'^2-2qq'+dq^2)]f(q'). \tag{2.8a}$$

Notice that $\exp(i\pi/4)C\left(\begin{smallmatrix}0&1\\-1&0\end{smallmatrix}\right)$ is the ordinary Fourier transform. When $|b|\to 0$, the integration kernel in (2.8) appears indeterminate, but can be shown to be well defined and turn (2.8) into

$$[C\left(\begin{smallmatrix}a&0\\c&a^{-1}\end{smallmatrix}\right)f](q)=a^{-1/2}\exp[(ic/2a)q^2]f(q/a). \tag{2.8b}$$

Formulas (2.8) give a unitary representation of $\mathrm{SL}(2,R)$ on $\mathcal{L}^2(R)$. This is actually a true representation of $\overline{\mathrm{SL}(2,R)}$, the covering group of $\mathrm{SL}(2,R)$ with respect to the O(2) subgroup generated by I_3; for $\mathrm{SL}(2,R)$ it is a ray representation and the possible phase differences

with a true representation have been discussed in Ref. 1.[22]

E. We can join the set of generators in w and sl(2,R) using the derivation property of the commutator bracket, and in the resulting algebra we find that w is an ideal. We thus define wsl(2,R) $= w \dot{+}$ sl(2,R), where $\dot{+}$ is the semidirect sum, as the collection of generators (2.1) and (2.5). Correspondingly, from W and SL(2,R) we build the semidirect product WSL(2,R) $= W \otimes$ SL(2,R) of pairs $g = \{\mathbf{A}, \omega\}$, and its unitary representation on $\mathcal{L}^2(R)$ is given by the composition of the constituent actions (2.3) and (2.8) as

$$[\mathcal{T}\{\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right), (xyz)\} f](q) \equiv [C\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) T_\omega(x,y,z) f](q)$$
$$= \int_R dq' B_g(q,q') f(q'), \tag{2.9a}$$

where the integral kernel $B_g(q,q')$ can be found[23] from (2.3) and (2.8); it will not be of interest by itself, indeed, the usefulness of the methods proposed in this article hinge upon our *not* needing the general form (2.9a), but only those transformations with $b = 0$ where the integral transform collapses to a *geometric* transform,

$$\mathcal{G}(a,c;x,y,z) \equiv \mathcal{T}\{\left(\begin{smallmatrix} a & 0 \\ c & a^{-1} \end{smallmatrix}\right), (x,y,z)\}, \tag{2.9b}$$

which has the effect

$$[\mathcal{G}(a,c;x,y,z) f](q)$$
$$= a^{-1/2} \exp[i(cq^2/2a + xq/a + \tfrac{1}{2}xy + z)] f(q/a + y), \tag{2.9c}$$

i.e., changes of scale (a), translations (y), multiplication by an exponential (x) and Gaussian (c), and an over-all phase (z). Notice that the composition of two geometric transforms is a geometric transform, and so is its inverse. Equation (2.9a) allows us, though, to write the WSL(2,R) product law for $g\{\mathbf{A}, \omega(x,y,z)\} \equiv g\{\mathbf{A}, \xi, z\}$ compactly as

$$g\{\mathbf{A}_1, \xi_1, z_1\} g\{\mathbf{A}_2, \xi_1, z_2\}$$
$$= g\{\mathbf{A}_1 \mathbf{A}_2, \ \xi_1 \mathbf{A}_2 + \xi_2, z_1 + z_2 + \tfrac{1}{2}\xi_1 \mathbf{A}_2 \Omega \xi_2^T\}, \tag{2.10}$$

so that the group indentity is $g\{\mathbf{1}, 0, 0\}$ and the inverse $g\{\mathbf{A}, \xi, z\}^{-1} = g\{\mathbf{A}^{-1}, -\xi\mathbf{A}^{-1}, -z\}$, where we have used the fact that $\mathbf{A}\Omega\mathbf{A}^T = \Omega$ and $\xi\Omega\xi^T = 0$ for $\mathbf{A} \in$ SL(2,R).

F. The action (2.9) of WSL(2,R) on $\mathcal{L}^2(R)$ induces its adjoint representation by automorphisms of the algebra,[4]

$$I_i \xrightarrow{g} I_i' = g I_i g^{-1} \equiv \mathrm{Ad}_g I_i = \sum_j M_{ij} I_j, \tag{2.11}$$

for $I_i \in$ wsl(2,R) denoting $I_4 = Q$, $I_5 = P$, and $I_6 = \mathbb{1}$. Through (2.1), (2.3), (2.5)–(2.7), and (2.9) we obtain[2]

$$\mathbf{M} = \begin{pmatrix} \frac{1}{2}(a^2 - b^2 - c^2 + d^2) & bd - ac & \frac{1}{2}(a^2 - b^2 + c^2 - d^2) & \frac{1}{2}(cx - dy) & \frac{1}{2}(-ax + by) & \frac{1}{4}(x^2 - y^2) \\ -ab + cd & ad + bc & -ab - cd & \frac{1}{2}(-dx - cy) & \frac{1}{2}(bx + ay) & -\frac{1}{2}xy \\ \frac{1}{2}(a^2 + b^2 - c^2 - d^2) & -ac - bd & \frac{1}{2}(a^2 + b^2 + c^2 + d^2) & \frac{1}{2}(cx + dy) & \frac{1}{2}(-ax - by) & \frac{1}{4}(x^2 + y^2) \\ & & & d & -b & y \\ & 0 & & -c & a & -x \\ & & & 0 & 0 & 1 \end{pmatrix}. \tag{2.12}$$

Since the group parameter z does not appear in (2.12), the latter is a faithful representation only of wsl(2,R)/$\mathbb{1}$. An operator H built as a linear combination of the generators of the algebra,

$$H = \sum_j \theta_j I_j = \tfrac{1}{2}(\theta_1 + \theta_3)P^2 + \tfrac{1}{4}\theta_2(QP + PQ) + \tfrac{1}{2}(-\theta_1 + \theta_3)Q^2$$
$$+ \theta_4 Q + \theta_5 P + \theta_6 \mathbb{1}, \tag{2.13a}$$

will transform under the adjoint action of the group as

$$H \xrightarrow{g} H' = gHg^{-1} = \sum_i \sum_j \theta_i M_{ij} I_j = \sum_j \theta_j' I_j. \tag{2.13b}$$

G. Two elements H and H' of the algebra are said to be on the same *orbit* under the group if there exists an element g in the group such that (2.13b) holds. Such elements H and H' generate one-parameter subgroups $g_0(\alpha) = \exp(i\alpha H)$ and $g_1(\beta) = \exp(i\beta H')$ which are conjugate through g, and thus $g_0(\alpha)$ and $g_1(\alpha)$ are in the same class in the group. Even if we perform an over-all change in scale $H'' = \gamma H'$ [which is *not* a transformation (2.12)–(2.13) for $|\gamma| \neq 1$], the subgroup generated as $g_2(\alpha)$ $= \exp(i\alpha H'') = g_1(\gamma\alpha) = g g_0(\gamma\alpha) g^{-1}$ will as *a whole* still be conjugate to the subgroup generated by H. Since the O(2) subgroup generated by $\mathbb{1}$ is a trivial phase, it will serve us to ignore it in our analysis, so that we will restrict our orbit analysis to the coset space[4] WSL(2,R)/O(2)$_{\mathbb{1}}$. In terms of the algebra wsl(2,R)/$\mathbb{1}$, this means that operators differing by an additive term $\theta_6\mathbb{1}$ are considered equivalent. In choosing the orbit representatives, over-all factors will also be disregarded since they generate the same subgroup.

H. The orbit structure of wsl(2,R)/$\mathbb{1}$ can now be analyzed,[14] noting that $\Theta \equiv \theta_3^2 - \theta_1^2 - \theta_2^2$ is an invariant under the transformation (2.13). As we are interested in operators equivalent up to over-all changes in scale γ (for which $\Theta'' = \gamma^2\Theta'$) we consider three cases: (i) $\Theta > 0$, (ii) $\Theta < 0$, and (iii) $\Theta = 0$. In each of these cases we can pick out an orbit *representative* operator H^ω, for each orbit ω. This is simplified by noting that we can choose the transformation to be a geometric transformation ($b = 0$) and that (2.12) has a lower-left zero submatrix.

(i) $\Theta > 0$ (*harmonic oscillator*):

$$H^h = 2I_3 = \tfrac{1}{2}(P^2 + Q^2), \tag{2.14a}$$

through

$$\begin{aligned} &a_h = [\theta_h/(\theta_1+\theta_3)]^{1/2}, \quad c_h = \theta_2[\theta_h(\theta_1+\theta_3)]^{-1/2}, \\ &x_h = 2\theta_h^{-2}[\theta_5(\theta_3-\theta_1) - \theta_4\theta_2], \\ &y_h = 2\theta_h^{-2}[\theta_5\theta_2 - \theta_4(\theta_3+\theta_1)], \end{aligned} \tag{2.14b}$$

where $\theta_h^2 = \Theta = \theta_3^2 - \theta_1^2 - \theta_2^2$ and the choice $\theta_h = 2$ leads to the form (2.14a). Clearly, the transformation (2.14b) is possible for all θ's except when $\theta_1 = -\theta_3$. This corresponds to the case when H has no P^2 (kinetic energy) term, which we can regard as unphysical. In this case, we can subject H to a Fourier transform, which is known and easy to implement, but is not a geometric transform.

(ii) $\Theta < 0$ (*repulsive oscillator*):

$$H^r = 2I_1 = \tfrac{1}{2}(P^2 - Q^2), \tag{2.15a}$$

through

$$\begin{aligned} &a_r = [\theta_r/(\theta_1+\theta_3)]^{1/2}, \quad c_r = \theta_2[\theta_r(\theta_1+\theta_3)]^{-1/2}, \\ &x_r = 2\theta_r^{-2}[\theta_5(\theta_1-\theta_3) + \theta_4\theta_2], \\ &y_r = 2\theta_r^{-2}[-\theta_5\theta_2 + \theta_4(\theta_3+\theta_1)], \end{aligned} \tag{2.15b}$$

where $\theta_r^2 = -\Theta = \theta_1^2 + \theta_2^2 - \theta_3^2$ and the choice $\theta_r = 2$ leads to (2.15a). Remarks as in (i) apply when $\theta_1 = -\theta_3$.

(iii) $\Theta = 0$ (*linear potential*):

$$H^l = I_1 + I_3 + Q = \tfrac{1}{2}P^2 + Q. \tag{2.16a}$$

Here we have several cases. As $\theta_1^2 + \theta_2^2 - \theta_3^2 = 0$ assume first θ_1, θ_2, and θ_3 are not all identically zero. Then through

$$\begin{aligned} &a_l = [2\theta_l/(\theta_1+\theta_3)]^{1/2}, \quad c_l = [(\theta_3-\theta_1)/2\theta_l]^{1/2}, \\ &x_l(\theta_3+\theta_1)^{1/2} - y_l(\theta_3-\theta_1)^{1/2} = 2\theta_5(\theta_3+\theta_1)^{-1/2} \end{aligned} \tag{2.16b}$$

we can bring H to the form H^l with θ_l a free parameter and $\theta_1' = \theta_l = \theta_3'$, while

$$\theta_4' = (2\theta_l)^{-1/2}[\theta_4(\theta_3+\theta_1)^{1/2} - \theta_5(\theta_3-\theta_1)^{1/2}]. \tag{2.16c}$$

The ratio $\rho \equiv \theta_4'/\theta_l$ can be varied by varying θ_l, and the choice (2.16a) corresponds to $\rho = 1$. We cannot make ρ vanish, however, unless to start with we have $\theta_4(\theta_3+\theta_1)^{1/2} = \theta_5(\theta_3-\theta_1)^{1/2}$. We distinguish this case:

(iii') $\Theta = 0$, $\theta_4^2(\theta_3+\theta_1) = \theta_5^2(\theta_3-\theta_1)$ (*free particle*):

$$H^f = I_1 + I_3 = \tfrac{1}{2}P^2, \tag{2.17}$$

and we must add the remark following (i) in the case $\theta_1 = -\theta_3$. Now we examine the cases where $\theta_1 = \theta_2 = \theta_3 = 0$. We only have the lower-right submatrix (2.12), and we can always bring the operator to the form

(iii'') $\theta_1 = 0$, $\theta_2 = 0$, $\theta_3 = 0$ (*momentum*):

$$H^m = P, \tag{2.18}$$

through

$$a_m = \theta_5^{-1}, \quad c_m = \theta_4, \tag{2.19}$$

applying the Fourier transformation when $\theta_5 = 0$. The further case when $\theta_4 = 0 = \theta_5$, has $\mathbf{0}$ for its orbit representative in $\mathrm{wsl}(2,R)/\mathbf{1}$ and $\mathbf{1}$ in $\mathrm{wsl}(2,R)$.

To sum up: We have five orbits in $\mathrm{WSL}(2,R)/\mathrm{O}(2)_1$ generated by H^ω ($\omega = h, r, l, f$ or m). We have found in each case the explicit transformation (2.12) leading a general operator (2.13a) to one of the five representatives, up to an over-all multiplicative constant and the (possible) addition of a multiple θ_6' of $\mathbf{1}$ given from (2.12)–(2.13) as

$$\begin{aligned} \theta_6' = {} & \tfrac{1}{4}(x_\omega^2 - y_\omega^2)\theta_1 - \tfrac{1}{2}x_\omega y_\omega\theta_2 + \tfrac{1}{4}(x_\omega^2 + y_\omega^2)\theta_3 \\ & + y_\omega\theta_4 - x_\omega\theta_5 + \theta_6 \end{aligned} \tag{2.20}$$

with x_ω, y_ω ($\omega = h, r, l, f$, or m) as in (2.14b), (2.15b), or (2.16b).

I. As the operators H as given by (2.13a) are self-adjoint in $\mathcal{L}^2(R)$, their eigenfunctions will constitute a complete orthonormal (possibly in the sense of Dirac) set of eigenvectors for the space, and since the transformations (2.9) are unitary, it suffices to give the results for the orbit representatives:

Harmonic Oscillator: These are well known[20] to be

$$\begin{aligned} &\psi_\lambda^h(q) = [2^n n!\sqrt{\pi}\,]^{-1/2}\exp(-\tfrac{1}{2}q^2)H_n(q), \\ &\lambda = n + \tfrac{1}{2}, \quad n = 0, 1, 2, \cdots, \end{aligned} \tag{2.21}$$

where $H_n(q)$ are the Hermite polynomials. Orthonormality has the usual phrasing as $(\psi_\lambda^h, \psi_{\lambda'}^h) = \delta_{\lambda,\lambda'}$ (Kronecker delta) and completeness states $\psi(q) = \sum \psi_\lambda^h(q)(\psi_\lambda^{r-}, \psi)$ in the norm for any $\psi \in \mathcal{L}^2(R)$.

Repulsive Oscillator: The basis and spectrum of $H^r = 2I_1$ can be found[14] in terms of that of $H^d \equiv -2I_2 = i(q\,d/dq + \frac{1}{2})$, which is on the same orbit: $H^r = g_{12}H^d g_{12}^{-1}$ with $g_{12} = \{(1/\sqrt{2})\left(\begin{smallmatrix}1 & 1\\ -1 & 1\end{smallmatrix}\right), (0)\}$ (this is the "square root" of the Fourier transform, as $g_{12}^2 = \{\left(\begin{smallmatrix}0 & 1\\ -1 & 0\end{smallmatrix}\right), (0)\}$). The eigenfunctions of H^d are found from the theory of Mellin transforms to be, properly normalized,

$$\psi_\lambda^{d\pm}(q) = (2\pi)^{-1/2} q_\pm^{-i\lambda-1/2}, \quad \lambda \in R, \quad q_\pm \equiv \begin{cases} \pm q, & q \gtrless 0, \\ 0, & q \lesseqgtr 0, \end{cases} \tag{2.22a}$$

with a spectrum covering twice the real line. Using (2.9a) for g_{12}, we can find $\psi_\lambda^r = \mathcal{F}(g_{12})\psi_\lambda^d$ as

$$\begin{aligned} \psi_\lambda^{r\pm}(q) = {} & 2^{-3/4}\,\pi^{-1}\exp[-(i/4)\pi(i\lambda + \tfrac{1}{2})] \\ & \times \Gamma(-i\lambda + \tfrac{1}{2})D_{i\lambda-1/2}(\pm 2^{1/2}\exp(3i\pi/4)q), \quad \lambda \in R, \end{aligned} \tag{2.22b}$$

where $D_\nu(r)$ is the Parabolic Cylinder function.[24] Orthonormality means here $(\psi_\lambda^{r\pm}, \psi_{\lambda'}^{r\pm}) = \delta(\lambda - \lambda')$ (Dirac delta) and $(\psi_\lambda^{r\pm}, \psi_{\lambda'}^{r\mp}) = 0$. Completeness integrates twice over $\lambda \in R$, i.e., $\psi(q) = \int_R d\lambda\,\psi_\lambda^{r+}(q)(\psi_\lambda^{r+}, \psi) + \int_R d\lambda\,\psi_\lambda^{r-}(q)(\psi^{r-}, \psi)$, $\psi \in \mathcal{L}^2(R)$.

Linear potential: Again, the basis and spectrum of H^l is easier to analyze[14] for its Fourier transform $\frac{1}{2}Q^2 - P$ which gives rise to a first-order differential equation whose normalized solutions are $\mathcal{U}_\lambda^l(q) = (2\pi)^{-1/2} \times \exp(-\lambda q + \frac{1}{6}q^3)$, for $\lambda \in R$. The inverse Fourier transform yields $\psi_\lambda^l(q)$ through Airy's integral[25]

$$\psi^I_\lambda(q) = 2^{1/3} Ai(2^{1/3}[q-\lambda]), \quad \lambda \in R, \tag{2.23}$$

and the usual orthonormality and completeness statements are $(\psi^I_\lambda, \psi^I_{\lambda'}) = \delta(\lambda - \lambda')$ and $\psi(q) = \int_R d\lambda\, \psi^I_\lambda(q)(\psi^I_\lambda, \psi)$, $\psi \in \mathcal{L}^2(R)$.

Free particle: The basis and the spectrum of P is

$$\psi^f_\lambda(q) = (2\pi)^{-1/2} \exp(i\lambda q), \quad \lambda \in R. \tag{2.24}$$

This serves also as a convenient basis for $H^f = \frac{1}{2}P^2$ which in linearly, but not functionally independent of P. The spectrum of H^f is $\frac{1}{2}\lambda^2$, i.e., twice the half-line.

The eigenfunctions ψ_μ and eigenvalues μ of an operator H as given by (2.13a) can now be determined, knowing the ones for the orbit representatives H^ω, ψ^ω_λ, and λ ($\omega = h, r, l$ or $f-m$). We have

$$g_\omega H g_\omega^{-1} = \theta_\omega H^\omega + \theta'_6 \mathbf{1} \tag{2.25a}$$

with g_ω a *geometric* transformation of the type (2.9b), with parameters given by (2.14b), (2.15b), or (2.16b) (save the cases when a Fourier transformation is needed) and the θ_ω determined correspondingly. Hence

$$\psi_\mu(q) = [\mathcal{G}(g_\omega^{-1})\psi^\omega_\lambda](q) \quad \text{and} \quad \mu = \theta_\omega \lambda + \theta'_6. \tag{2.25b}$$

Recall that geometric transforms are easily obtained as in (2.9c).

J. Example:

$$\begin{aligned} H &= 2P^2 + (QP+PQ) + \tfrac{1}{2}Q^2 + Q + P + \xi\mathbf{1} \\ &= 3I_1 + 4I_2 + 5I_3 + Q + P + \xi\mathbf{1}. \end{aligned} \tag{2.26a}$$

We see that $\Theta = 0$, so this case belongs to (iii). From (2.16b) we find $a_l = \frac{1}{2}\sqrt{\theta_l}$, $c_l = 1/\sqrt{\theta_l}$ and $2x_l - y_l = \frac{1}{2}$. The transformation

$$\left\{ \begin{pmatrix} \frac{1}{2}\sqrt{\theta_l} & 0 \\ 1/\sqrt{\theta_l} & 2/\sqrt{\theta_l} \end{pmatrix}, (x_l, 2x_l - \tfrac{1}{2}, 0) \right\}$$

then maps H into $H' = \frac{1}{2}\theta_l P^2 + 1/\sqrt{\theta_l}\, Q + (x_l + \xi - 3/8)\mathbf{1}$ so we choose $\theta_l = 1$ and $x_l = \frac{3}{8} - \xi$. The spectrum of H is then $\mu = \lambda \in R$, while the basis functions are

$$\begin{aligned} &\psi_\lambda(q) \\ &= [\mathcal{G}(\tfrac{1}{2}, 1; \tfrac{3}{8} - \xi, \tfrac{1}{4} - 2\xi, 0)^{-1} \psi^I_\lambda](q) \\ &= [\mathcal{G}(2, -1; -\tfrac{1}{2}, \xi - \tfrac{1}{8}, 0)\psi^I_\lambda](q) \\ &= 2^{-1/2} \exp i[-\tfrac{1}{4}(q^2 + q + \xi - 1/8)]\, \psi^I_\lambda(\tfrac{1}{2}q + \xi - \tfrac{1}{8}) \\ &= 2^{-1/6} \exp i[-\tfrac{1}{4}(q^2 + q + \xi - 1/8)]\, \mathrm{Ai}(2^{1/3}\{\tfrac{1}{2}q + \xi - \tfrac{1}{8} + \lambda\}). \end{aligned} \tag{2.26b}$$

III. COMPLEX CANONICAL TRANSFORMS AND TIME DEVELOPMENT OF A SYSTEM

A. We will now allow the group parameters of $g = \{\mathbf{A}, \xi, z\} \in \mathrm{WSL}(2,R)$ (det $\mathbf{A} = 1$) to range over the complex field. The resulting set also forms a group which we denote by $\mathrm{WSL}(2,C)$. The representation given by (2.3)—(2.8) and (2.9) does not follow for the whole new group: If f is assumed to be in $\mathcal{L}^2(R)$, $\mathcal{J}f$ will belong to $\mathcal{L}^2(R)$ only if the kernel $B_g(q,q')$ is bounded. This happens for the parameters in $\mathbf{A}$ only if $\mathrm{Im}(a/b) \geq 0$ so that the Gaussian factor will be decreasing and, when $a = 0$, b must be real so that the kernel will be an oscillating exponential. For the $\omega(x,y,z)$ parameters it is only required that when $a = 0$, x be real also. The product of two bounded operators is bounded and the group identity is bounded as well as all real elements in $\mathrm{WSL}(2,R)$. Thus, (2.9a) represents properly a *subsemigroup* of $\mathrm{WSL}(2,C)$ which we denote by $\mathrm{HWSL}(2,C)$, following Refs. 1, 2, and 7 which deal with the $\mathrm{SL}(2,C)$ part. As regards unitarity, those transformations in $\mathrm{HWSL}(2,C)$ which are *not* in $\mathrm{WSL}(2,R)$, are represented by integral nonunitary transformations from $\mathcal{L}^2(R)$ into $\mathcal{L}^2(R)$.

B. In Refs. 1 and 2, we constructed Hilbert spaces of analytic functions $\mathcal{J}_\mathbf{A}$ such that $\mathrm{HSL}(2,C)$ is represented by *unitary* mappings between $\mathcal{L}^2(R)$ and $\mathcal{J}_\mathbf{A}$. The Hilbert spaces $\mathcal{J}_\mathbf{A}$ are characterized by a scalar product performed over the complex plane, as in Bargmann's case,[8] given by

$$(f,g)_\mathbf{A} = \int_\mathbb{C} d^2\mu(q)\, f(q)^*\, g(q), \tag{3.1a}$$

with the measure

$$\begin{aligned} d^2\mu(q) &= 2(2\pi v)^{-1/2} \exp[(1/2v)(uq^2 - 2qq^* + u^*q^{*2})] \\ &\quad \times d\,\mathrm{Re}q\, d\,\mathrm{Im}q, \end{aligned} \tag{3.1b}$$

and where

$$u = a^*d - b^*c, \quad v = 2\,\mathrm{Im}b^*a > 0. \tag{3.1c}$$

Corresponding to the geometric transformations (2.9b), where $v = 0$ the measure becomes singular and one can show that

$$\lim_{v\to 0} \int_\mathbb{C} d^2\mu(q)\, f(q)^*\, g(q) = \int_{Re^{i\psi}} dx \exp(-w|x|^2/2) f(x)^*\, g(x), \tag{3.1d}$$

where $w = 2\,\mathrm{Im}c^*d$, and the integration contour is along a line in the complex $\mathbb{C}$-plane tilted with respect to the real axis by an angle $\psi = -\frac{1}{2}$ phase of u. Finally, for the general complex case, the transform inverse to (2.8) is given by

$$f(q) = \int_\mathbb{C} d^2\mu(q') A(q',q)^*\, [\mathcal{C}f](q'). \tag{3.1e}$$

With little extra labor we can build a similar scalar product and Hilbert spaces such that the transformations in $\mathrm{HWSL}(2,C)$ will be unitary. The only application we will touch upon is the one provided by the heat equation, and so the construction of the general case beyond (3.1) is unnecessary here.

C. The action of $\mathrm{WSL}(2,C)$ transformations on operators H of the form (2.13a) closely follows that seen in the last section, except for allowing all parameters to be complex. The orbit structure analyzed in II-H simplifies, in that the cases (i) and (ii) (attractive and repulsive oscillators) coalesce, if we allow for over-all complex factors. Indeed, the well-known Bargmann transformation,[8] $g_B \equiv \{(1/\sqrt{2})\begin{pmatrix} 1 & -i \\ -i & 1 \end{pmatrix}, (0)\}$, bridges (i) and (ii), as $g_B I_3 g_B^{-1} = iI_2$ while $g_H \equiv \{\begin{pmatrix} \omega & 0 \\ 0 & \omega^{-1} \end{pmatrix}, (0)\}$, $\omega^2 = -i$, performs $g_H H^f g_H^{-1} = -iH^f$ and takes us from the free particle Schrödinger equation to the heat equation.

D. The parabolic differential equations we want to analyze here are those of the general form

$$Hu(q,t) = -i(\partial/\partial t)u(q,t), \tag{3.2}$$

where H is an operator of the form (1.6a)— (2.13a).

Formally, the solution of (3.2) is given by the time-translated initial condition $u(q) \equiv u(q,0)$,

$$u(q,t) = \exp(t\,\partial/\partial t')\,u(q,t')\,|_{t'=0}$$
$$= \exp(itH)\,u(q) \equiv [H_t u](q). \tag{3.3}$$

The third term in (3.3) is a differential operator of infinite degree in q (termed also hyperdifferential operator[26]) densely defined in $\mathcal{L}^2(R)$, whose action on $u(q)$ is a time-dependent canonical transform H_t whose integral form is given by (2.9). Corresponding to the four orbits seen in the last section (excluding P), their four H_t^w time-evolution operators are represented by

$$H_t^h = \exp(it\tfrac{1}{2}[P^2+Q^2]):\left\{\begin{pmatrix}\cos t & -\sin t\\ \sin t & \cos t\end{pmatrix},\ (0,0,0)\right\} \tag{3.4a}$$

$$H_t^r = \exp(it\tfrac{1}{2}[P^2-Q^2]):\left\{\begin{pmatrix}\cosh t & -\sinh t\\ -\sinh t & \cosh t\end{pmatrix},\ (0,0,0)\right\} \tag{3.4b}$$

$$H_t^l = \exp(it[\tfrac{1}{2}P^2+Q]):\left\{\begin{pmatrix}1 & -t\\ 0 & 1\end{pmatrix},\ (-t,\tfrac{1}{2}t^2,-(1/6)t^3)\right\} \tag{3.4c}$$

$$H_t^f = \exp(it\tfrac{1}{2}P^2):\left\{\begin{pmatrix}1 & -t\\ 0 & 1\end{pmatrix},\ (0,0,0)\right\}. \tag{3.4d}$$

All these expressions can be read off from (2.7), except for (3.4c), which requires some extra work in exponentiating.

For a general operator H [(2.13a)] we can find its geometric transformation g_ω ($\omega = h,r,l$ or f) relating it to its orbit representative. Its time-evolution transform will be

$$H_t = \exp(itH) = \exp(it\theta_6')\exp(it\theta_\omega g_\omega^{-1} H^\omega g_\omega)$$
$$= \exp(it\theta_6')\,g_\omega^{-1} H^\omega_{\theta_\omega t}\,g_\omega, \tag{3.5a}$$

and its solutions

$$u(q,t) = H_t u(q) = \exp(it\theta_6')\,\mathcal{G}(g_\omega^{-1}) H^\omega_{\theta_\omega t}\,\mathcal{G}(g_\omega)u(q). \tag{3.5b}$$

E. Simplest to consider, is the time evolution of the eigenfunctions $\psi_\lambda(q)$ of the operator H in (3.2), since

$$\psi_\lambda(q,t) = H_t\psi_\lambda(q) = \exp(i\lambda t)\,\psi_\lambda(q). \tag{3.6}$$

These are the solutions of (3.2) *separable* in q and t: if we know the expansion coefficients, u_λ of an arbitrary function $u(q) = u(q,0)$ in terms of the ψ_λ-basis, the expansion coefficients of the $u(q,t)$ solution of (3.2) are $u_\lambda \exp(i\lambda t)$. But assume that the physically meaningful expansion for $u(q)$ is in terms of a $\psi'_\lambda(q)$-basis, eigenfunctions of an operator H' which may or may not be on the same orbit as H. Assume for definiteness that H and H' are the orbit representatives of the last section, with (3.4) their time-evolution transforms. Then, it is fundamental for our results that, at least in a region around $t=0$, we can write

$$H_t = \mathcal{G}_t H'_{t'}, \tag{3.7}$$

where $t' = t'(t)$. That is, the time-evolution transform H_t can be written as the time-evolution transform $H'_{t'}$ for a rescaled time $t'(t)$, times a (time-dependent) *geometric* transform $\mathcal{G}_t$. Finding the group parameters of $\mathcal{G}_t$ and the function $t'(t)$ is an exercise in 2×2 matrix algebra.

F. Example: Let H be the harmonic oscillator Schrödinger Hamiltonian [so that H_t is H_t^h in (3.4a)]. We want to find the time evolution of plane waves [free particle eigenfunctions ψ_λ^f in (2.24), $H'_{t'}$ being $H_{t'}^f$] in that system. We write (3.7), where only the SL(2,R) parameters need to be considered, as

$$\begin{pmatrix}\cos t & -\sin t\\ \sin t & \cos t\end{pmatrix} = \begin{pmatrix}a_t & 0\\ c_t & a_t^{-1}\end{pmatrix}\begin{pmatrix}1 & -t'\\ 0 & 1\end{pmatrix}, \tag{3.8a}$$

and we find immediately,

$$t' = \tan t,\quad a_t = \cos t,\quad c_t = \sin t, \tag{3.8b}$$

so, from (3.6) and (2.9),

$$\psi_\lambda^f(q) \xrightarrow{H_t^h} \psi_\lambda^f(q) = \mathcal{G}_t\,H_{t'}^f\,\psi_\lambda^f(q)$$
$$= \exp(i\tfrac{1}{2}\lambda^2 t')\,\mathcal{G}_t\psi_\lambda^f(q)$$
$$= \exp(i\tfrac{1}{2}\lambda^2 t')\,a_t^{-1/2}\exp(ic_t q^2/2a_t)\psi_\lambda^f(q/a_t)$$
$$= (\cos t)^{-1/2}\exp[i\tan t(\tfrac{1}{2}\lambda^2+\tfrac{1}{2}q^2)]\psi_\lambda^f(q/\cos t)$$
$$= \exp[i\tfrac{1}{2}\sin t\cos t(q/\cos t)^2]\psi_\lambda^f(q/\cos t)$$
$$\times(\cos t)^{-1/2}\exp(i\tfrac{1}{2}\lambda^2\tan t). \tag{3.9c}$$

This result can be checked using the harmonic oscillator Green's function and performing the integration.

A few comments on (3.9c): Although the points $t = \pm\frac{1}{2}\pi,\ \pm(3/2)\pi,\ \cdots$ appear to be singular for some elements of the expression, since the transformation (3.7) is unitary in $\mathcal{L}^2(R)$, we are assured that any $\mathcal{L}^2(R)$ function expanded in the ψ_λ^f-basis will exhibit no singularities in its time development. Systems which classically are periodic or exhibit turning points will be in many-to-one correspondence with open systems. In Table I we give, for all pairs of orbit representatives, the geometric transformation which bridges them.

G. The next point to be remarked upon is that the final expression in (3.9c) is (from right to left) a product of a function in $t'(t)$ times a function in $v(q,t) = q/\cos t$ times a *multiplier* $\exp(i\frac{1}{4}v^2\sin 2t)$. If we follow the procedure of Kalnins, Miller, and Boyer[14,15] in finding coordinate systems $v(q,t) - t$ such that, in (3.2),

$$u_\lambda(q,t) = \exp[iS(v,t)]\,V_\lambda(v)\,T_\lambda(t) \tag{3.10}$$

separates into two ordinary differential equations in v and t, one of such systems will be the one found above. The presence of the exponent in $S(v,t)$ (specifically *not* a sum of a function in v plus a function in t), defines this case as *R-separable*, as opposed to ordinary separability, when $S(v,t) = 0$. It is thus that, as detailed below, we obtain all "separating" coordinate systems for all parabolic equations (3.2). We follow the procedure of the example in subsection III. F to read them off Table I as

TABLE I. Expressions for the geometric transformations between pairs of time-development operators corresponding to the four orbit representatives $H_t^{\omega}=\mathcal{G}_t(a,c;x,y,z)H_{t'}^{\omega'}$. The entry "1" means $t=t'$ and $\mathcal{G}$ is the identity transformation. When x, y, z do not appear, they equal 0. The example in Sec. II. E corresponds to $\omega=h, \omega'=f$. The heat equation follows the f-system with the replacement $t \to 2it$.

ω \ ω'	h	r	l	f
h	1	$\tanh t' = \tan t$ $a_t = (\cos 2t)^{1/2}$ $c_t = \sin 2t (\cos 2t)^{-1/2}$	$t' = \tan t$ $a_t = \cos t$ $c_t = \sin t$ $x = t', y = \frac{1}{2}t'^2$ $z = \frac{1}{6}t'^3$	$t' = \tan t$ $a_t = \cos t$ $c_t = \sin t$
r	$\tan t' = \tanh t$ $a_t = (\cosh 2t)^{1/2}$ $c_t = \sinh 2t (\cosh 2t)^{-1/2}$	1	$t' = \tanh t$ $a_t = \cosh t$ $c_t = \sinh t$ $x = t', y = \frac{1}{2}t'^2$ $z = \frac{1}{6}t'^3$	$t' = \tanh t$ $a_t = \cosh t$ $c_t = \sinh t$
l	$\tan t' = t$ $a_t = (1+t^2)^{1/2}$ $c_t = t(1+t^2)^{-1/2}$ $x_t = -t(1+\frac{1}{2}t^2)(1+t^2)^{-1/2}$ $y_t = -\frac{1}{2}t^2(1+t^2)^{-1/2}$ $z_t = -\frac{1}{6}t^3$	$\tanh t' = t$ $a_t = (1-t^2)^{1/2}$ $c_t = t(1-t^2)^{-1/2}$ $x_t = -t(1-\frac{1}{2}t^2)(1-t^2)^{-1/2}$ $y_t = -\frac{1}{2}t^2(1-t^2)^{-1/2}$ $z_t = -\frac{1}{6}t^3$	1	$t' = t$ $a_t = 1, \ c_t = 0$ $x_t = -t$ $y_t = -\frac{1}{2}t^2$ $z_t = -\frac{1}{6}t^3$
f	$\tan t' = t$ $a_t = (1+t^2)^{1/2}$ $c_t = -t(1+t^2)^{-1/2}$	$\tanh t' = t$ $a_t = (1-t^2)^{1/2}$ $c_t = t(1-t^2)^{-1/2}$	$t' = t$ $a_t = 1, \ c_t = 0$ $x_t = t, \ y_t = \frac{1}{2}t^2$ $z_t = \frac{1}{6}t^3$	1

follows: From (2.9c) and (3.7),

$$
\begin{aligned}
&H_t^{\omega}\psi_{\lambda}^{\omega'}(q) \\
&\quad = \mathcal{G}_t H_{t'}^{\omega'}\psi_{\lambda}^{\omega'}(q) = \exp(i\lambda t')\mathcal{G}_t\,\psi_{\lambda}^{\omega'}(q) \\
&\quad = (a_t)^{-1/2}\exp\{i[c_t/2a_t)q^2 + (x_t/a_t)q + z_t + \tfrac{1}{2}x_t y_t + \lambda t']\} \\
&\qquad \times \psi_{\lambda}^{\omega'}(q/a_t + y_t) \\
&\quad = (a_t)^{-1/2}\exp\{i[\tfrac{1}{2}c_t a_t v^2 + (v - \tfrac{1}{2}y_t)(x_t - c_t a_t y_t) \\
&\qquad + z_t + \lambda t']\}\psi_{\lambda}^{\omega'}(v),
\end{aligned}
$$

where

$$
v(q,t) = q/a_t + y_t, \tag{3.11b}
$$

and all other parameters in $\mathcal{G}_t$, a_t, b_t, ..., z_t, and t' depend on t only. Thus $H_t^{\omega}\psi_{\lambda}^{\omega'}(q)$ is a separable function in the sense (3.10) in v and t, where the multiplier $S(v,t)$ can be read off (3.11a) and is

$$
S(v,t) = \tfrac{1}{2}c_t a_t v^2 + (x_t - c_t a_t y_t)v, \tag{3.11c}
$$

where as stated, a_t, c_t, x_t, y_t depend on t.

The differential equation (3.2) for H^{ω} generating H_t^{ω} will separate in two differential equations, one of the form of an eigenvalue equation for $H^{\omega'}$ in the variable $v(q,t)$ and the other, a first-order differential equation in t. This can be seen by writing (3.7) for $t \to 0$, as $\partial t'/\partial t|_{t=0} = 1$; it yields

$$
H^{\omega} = G + H^{\omega'}, \tag{3.12}
$$

where G generates $\mathcal{G}_t$ and is thus a first-order differential operator in q. The part in the separable function which depends only on v is $\psi_{\lambda}^{\omega'}$, which was chosen as an eigenfunction of $H^{\omega'}$ to start with. We have used $H^{\omega'}$ to separate the variables for H^{ω} in (3.2).

We can now see *a posteriori* why the factorization (3.7) works for all orbit representatives: They all have the form $\frac{1}{2}P^2 + V(Q)$ so that G will only be a function of Q. A disentanglement of the Baker—Campbell—Hausdorff type to produce (3.7) out of (3.12) will introduce the P and $PQ+QP$ parts which generate the translations and scale transformation. In Table II we have collected the separating coordinates and multipliers for all pairs of orbit representatives. The results can be compared with the literature.[27] In order to describe the general form of the separating coordinates and to determine the H' to which they correspond, defining equivalence between coordinate pairs, we must present first the material of the next section. The general case, however, can be formulated as follows.

H. We are given arbitrary H and H', and we can determine the (geometric) transformation relating them to their orbit representatives. We are thus able to know their time-evolution transforms H_t and H'_t through (3.5). We can write $H_t = \{\left(\begin{smallmatrix} h_a & h_b \\ h_c & h_d \end{smallmatrix}\right), (h_x, h_y, h_z)\}$ with $h_i = h_i(t)$, $(i = a, b, \ldots, z)$, where, it should be noticed, the $h_i(t)$ are linear combinations of trigonometric, hyperbolic, or power functions of t when H lies in the h, r, or l—f orbits. A similar construction is done for $H'_{t'}$ with $h'_i(t')$, and the product with a general $\mathcal{G}_t$ is made as in (3.7). Comparison of the ratio of the 1—1 and 1—2 matrix elements gives

TABLE II. Expressions for the coordinate systems $(v(q,t),t)$ which separate the equation $H^{\omega}\psi = -i\partial_t\psi$ into two ordinary differential equations in v and t, such that $\psi(q,t) = e^{iS(v,t)}V(v)T(t)$. The separation operator is $H^{\omega'}$. The heat equation follows the f-case with $t \to 2it$.

ω \ ω'	h	r	l	f
h	$v = q$ $S = 0$	$v = q(\cos 2t)^{-1/2}$ $S = \frac{1}{2}v^2 \sin 2t$	$v = q/\cos t + \frac{1}{2}\tan^2 t$ $S = \frac{1}{4}v^2 \sin 2t$ $+ v \tan t(1 - \frac{1}{2}\sin^2 t)$	$v = q/\cos t$ $S = \frac{1}{4}v^2 \sin 2t$
r	$v = q(\cosh 2t)^{-1/2}$ $S = \frac{1}{2}v^2 \sinh 2t$	$v = q$ $S = 0$	$v = q/\cosh t + \frac{1}{2}\tanh^2 t$ $S = -\frac{1}{4}v^2 \sinh 2t$ $+ v \tanh t(1 + \frac{1}{2}\sinh^2 t)$	$v = q/\cosh t$ $S = -\frac{1}{4}v^2 \sinh 2t$
l	$v = q(1+t^2)^{-1/2}$ $S = \frac{1}{2}v^2 t - vt(1+t^2)^{1/2}$	$v = q(1-t^2)^{-1/2}$ $S = \frac{1}{2}v^2 t - vt(1-t^2)^{-1/2}$	$v = q$ $S = 0$	$v = q - \frac{1}{2}t^2$ $S = -vt$
f	$v = q(1+t^2)^{-1/2}$ $S = -\frac{1}{2}v^2 t$	$v = q(1-t^2)^{-1/2}$ $S = \frac{1}{2}v^2 t$	$v = q + \frac{1}{2}t^2$ $S = vt$	$v = q$ $S = 0$

$$h(t) \equiv h_a(t)/h_b(t) = h'_a(t')/h'_b(t') \equiv h'(t'), \tag{3.13a}$$

whereby all h'_i's are known as functions of t. This is valid whenever h_b and h'_b are different from zero (this is *not* the case when H or H' is θI_2, for example). The parameters in the geometric transformation are then found as

$$\begin{aligned} a_t &= h_a/h'_a = h_b/h'_b, \\ c_t &= h_c/h'_a - h'_c/h_a = h_d/h'_b - h'_d/h_b, \\ (x_t, y_t) &= (h_x - h'_x, h_y - h'_y)\begin{pmatrix} h'_a & -h'_b \\ -h'_c & h'_a \end{pmatrix}, \end{aligned} \tag{3.13b}$$

and the separating variables and multipliers are found as in (3.11b)–(3.11c).

I. These developments also apply to complex transforms. Of particular interest is the heat equation,

$$\frac{\partial^2}{\partial q^2}u(q,t) = \frac{\partial}{\partial t}u(q,t), \tag{3.14a}$$

i.e., in the form (3.2), $H^H = 2iH^f$. In the form (2.25a) this corresponds to $\theta_f = 2$, $\theta'_6 = 0$ and a scale transformation with $a^2 = i$ (subsection III.C). Better still, we can set $\theta_f = 2i$ and Eqs. (3.4d)–(3.5a) then state that the time-evolution transform is,

$$\mathcal{H}^H_t = \mathcal{H}^f_{2it} = \exp\left(t\frac{\partial^2}{\partial q^2}\right) : \left\{ \begin{pmatrix} 1 & -2it \\ 0 & 1 \end{pmatrix}, (0) \right\}. \tag{3.14b}$$

The separable solutions, coordinates, and multipliers for the heat equation, with respect to each of the orbit representatives we have considered, can thus be read off the bottom row of Table II, replacing $t \to 2it$. We have thus the separable solutions in terms of oscillator, parabolic cylinder, Airy, and exponential functions.[28]

J. In comparing with the literature,[29] we notice that one of the better-known separating coordinate systems, that giving rise to the heat polynomials[30] $v_n(q,t) \equiv (-t)^{n/2}H_n(\frac{1}{2}q[-t]^{-1/2})$, solutions of (3.14a), is apparently missing. We proceed to show that it is related to the entry in the h-orbit.

The Hermite differential equation can be written as

$$\begin{aligned} DH_n(q) &\equiv \left(-\frac{1}{2}\frac{d^2}{dq^2} + q\frac{d}{dq} + \frac{1}{2}\right)H_n(q) \\ &= (I_1 + 2iI_2 + I_3)H_n(q) = (n + \tfrac{1}{2})H_n(q), \end{aligned} \tag{3.15a}$$

so that $\Theta = 4 > 0$ and we can write $g_h D g_h^{-1} = \theta_h I_3 = \frac{1}{2}\theta_h H^h$ finding g_h to be a geometric SL(2,C) transformation given by (2.14) with $\theta_h = 2$, $a_h = 1$, $c_h = i$. This is a complex canonical transform, so that the eigenfunctions of D, $H_n(q)$, will be orthogonal with respect to the measure given by (3.1d) which is $e^{-q^2}dq$ and the integration performed over the real line[31] as in (3.1d). The time-development operator is

$$D_{t'} = g_h^{-1}\mathcal{H}^h_{t'}g_h : \left\{ \begin{pmatrix} \exp(-it') & -\sin t' \\ 0 & \exp(it') \end{pmatrix}, (0) \right\} \tag{3.15b}$$

and the decomposition $\mathcal{H}^H_t = \mathcal{G}_t D_{t'}$ is possible with $a = \exp(it') = (1-4t)^{1/2}$, $c_t = 0$. This yields

$$\begin{aligned} \mathcal{H}^H_t H_n(q) &= \exp[i(n + 1/2)t']\mathcal{G}_t H_n(q) \\ &= (1-4t)^{n/2}H_n(q[1-4t]^{-1/2}) = 2^n v_n(q, t - \tfrac{1}{4}), \end{aligned} \tag{3.15c}$$

which is a polynomial in q and $t - \frac{1}{4}$.

The separating coordinates are $v = q(1-4t)^{-1/2}$ and t equivalent under time translation to $\frac{1}{2}q(-t)^{-1/2}$, t and the multiplier $S(v,t)$ is zero. From (3.15c) we see that if the temperature distribution of a conducting rod at $t = 0$ is $H_n(q) = 2^n v_n(q, -\frac{1}{4})$, it will evolve in time as $2^n v_n(q, t - \frac{1}{4})$ and at $t = \frac{1}{4}$, $2^n v_n(q, 0) = (2q)^n$. It should be observed that the $v_n(q, t - \frac{1}{4})$ are not elements of $\mathcal{L}^2(R)$ [nor is D self-adjoint in $\mathcal{L}^2(R)$]. However, as remarked above, D *is* self-adjoint if we take the measure $e^{-q^2}dq$, and there its eigenvectors are orthogonal and complete. Were we looking for the separating operator which produces the heat polynomials themselves, as $v_n(q,0) = q^n$, the operator would have been $H' \sim iI_2$. For this operator, however, we have $h'_b = 0$ and the decomposition (3.7) fails.

It should be observed that, since $H^H = -i\,\partial^2/\partial q^2$ is not Hermitian in $\mathcal{L}^2(R)$, the time development operator for the solutions of the heat equation (3.14b) is not unitary and does not preserve the orthogonality of two functions $f(q,t)$, $g(q,t)$ in $\mathcal{L}^2(R)$. However, if we use the formalism of complex canonical transforms, $\mathcal{H}_t^H$ is made a unitary mapping between $\mathcal{L}^2(R) \equiv \mathcal{J}_0$ and spaces $\mathcal{J}_t$ where the scalar product is, from (3.14b) and (3.1),

$$(f(\circ,t),\, g(\circ,t))_t \equiv \int_{\mathbb{C}} d\,\mathrm{Re}q\, d\,\mathrm{Im}q\, (2\pi t)^{-1/2} \exp[-(\mathrm{Im}q)^2/t]\, f(q,t)^*\, g(q,t), \quad t \geq 0. \tag{3.16}$$

Thus we can state that the quantity (3.16) is a *quadratic invariant* of the heat equation under time translations. This invariant is distinct from the total heat (a linear invariant), and is apparently new. Indeed, any differential equation (3.2) of the type we are studying will have its corresponding quadratic invariant.

IV. INVARIANCE GROUP AND INVARIANT BOUNDARIES

A. Lie theory has been used to solve partial differential equations through exploring their invariance under infinitesimal transformations, reducing thus by one the number of variables and then determining the subgroup—which leaves invariant a particular set of boundary conditions.[12] These methods apply to linear or nonlinear equations of any order. By contrast, our procedure is designed for linear parabolic equations of the type (1.6)—(3.2) and solves the problem through the use of matrix algebra in a global rather than infinitesimal manner.

The *invariance* of (3.2) under a transformation $g \in \mathrm{WSL}(2,C)$ can be stated as follows: when $u(q,t)$ is a solution of (3.2), then $v(q,t) \equiv \mathcal{J}_g^{(t)} U(q,t)$, where $\mathcal{J}_g^{(t)}$ is a two-variable representation of a canonical transform, is also a solution of (3.2). Notice that we have not said "if": *Any* such function will be a solution and the full invariance group of the equation will be the group WSL(2, C of six (complex) parameters. We will show below that, moreover, $v(q,t)$ will have the form

$$v(q,t) = \mathcal{J}_g^{(t)} u(q,t) = \mu_g(q,t)\, u(\bar{q}_g(q,t),\, \bar{t}_g(t)), \tag{4.1}$$

where μ_g, $\bar{q}_g$, and $\bar{t}_g$ are determinable functions of q and t. We should impose the additional conditions, however, that if q and t are real, then $\bar{q}_g$ and $\bar{t}_g$ should be also real and that if u is either square-integrable or real (the latter case in the heat equation, for example), then so should (4.1) be. This will reduce the acceptable symmetry group to a real subgroup of WSL(2, C).

B. In order to prove (4.1) and find the functions involved, use (2.9), (3.2)—(3.3), and (3.11): if $u(q,t)$ is the time development of the initial conditions $u(q) \equiv u(q,0)$ then $v(q,t) = \mathcal{J}_g^{(t)} u(q,t)$ is the time development of $v(q) = (\mathcal{J}_g u)(q)$:

$$\begin{aligned} v(q,t)) &= (\mathcal{J}_g^{(t)} u)(q,t) = (\mathcal{H}_t v)(q) \\ &= (\mathcal{H}_t \mathcal{J}_g u)(q) = (\mathcal{G}_{\bar{g}(g,t)}\, \mathcal{H}_{\bar{t}_g(t)}\, u)(q) \\ &= (\mathcal{G}_{\bar{g}} u)(q, \bar{t}_g) \\ &= \bar{a}^{-1/2} \exp\{i[(\bar{c}/2\bar{a})q^2 + (\bar{x}/\bar{a})q + \tfrac{1}{2}\bar{x}\bar{y} + \bar{z}]\} \\ &\quad \times u(\bar{q}_g(q,t),\, \bar{t}_g(t)), \end{aligned} \tag{4.2a}$$

where $\bar{a} = \bar{a}(t), \ldots, \bar{z} = \bar{z}(t)$ and

$$\bar{q}_g(q,t) = (q/\bar{a}) + \bar{y}, \quad h(\bar{t}_g) = [dh(t) + b]/[a + ch(t)] \tag{4.2b}$$

with the function $h(t)$ defined as in (3.13a). The key step in (4.2a) has been that of writing $\mathcal{H}_t \mathcal{J}_g = \mathcal{G}_{\bar{g}} \mathcal{H}_{\bar{t}}$, i.e., time development × canonical transform = geometric transform × time development in $\bar{t}_g(t)$. The last member of (4.2a) and (4.2b) were obtained from (3.11a)—(3.11b).

C. As a first illustration of (4.2) consider the case of the free particle, closely related to the heat equation, where the results are known[12,14]:

$$\begin{aligned} \mathcal{H}_t^f \mathcal{J}_g &= \left\{\begin{pmatrix} 1 & -t \\ 0 & 0 \end{pmatrix},\, (0)\right\} \left\{\begin{pmatrix} a & b \\ c & d \end{pmatrix},\, (xyz)\right\} \\ &\subset \left\{\begin{pmatrix} a - ct & b - dt \\ c & d \end{pmatrix},\quad (x,y,z)\right\} \\ &= \mathcal{G}_{\bar{g}} \mathcal{H}_{\bar{t}}^f = \left\{\begin{pmatrix} \bar{a} & 0 \\ \bar{c} & \bar{a}^{-1} \end{pmatrix} (\bar{x}\bar{y}\bar{z})\right\} \left\{\begin{pmatrix} 1 & -\bar{t} \\ 0 & 1 \end{pmatrix},\, (0)\right\} \\ &= \left\{\begin{pmatrix} \bar{a} & -\bar{a}\bar{t} \\ \bar{c} & -\bar{c}\bar{t} + \bar{a}^{-1} \end{pmatrix},\, (\bar{x}, \bar{y} - \bar{x}\bar{t}, \bar{z})\right\}. \end{aligned} \tag{4.3a}$$

Equation (4.3a) contains six independent simultaneous equations which yield

$$\bar{a} = a - ct, \quad \bar{c} = c, \quad \bar{x} = x, \quad \bar{y} = y + x\bar{t}, \quad \bar{z} = z \tag{4.3b}$$

and from (4.2b)

$$\begin{aligned} \bar{q} &\equiv \bar{q}_g(q,t) = (q + xdt - xb)/(a - ct) + y, \\ \bar{t} &\equiv \bar{t}_g(t) = (dt - b)/(a - ct). \end{aligned} \tag{4.3c}$$

Hence, if $u(q,t)$ is a solution of the free-particle Schrödinger equation, then so is

$$\begin{aligned} V(q,t) &= \mathcal{G}_{\bar{g}} u(q,\bar{t}) \\ &= (a - ct)^{-1/2} \\ &\quad \exp(i\{(a - ct)^{-1}[cq^2 + xq + \tfrac{1}{2}x^2(dt - b)] \\ &\quad + \tfrac{1}{2}xy + z\}) \\ &\quad u((q + xdt - xb)/(a - ct) \\ &\quad + y, (dt - b)/(a - ct)). \end{aligned} \tag{4.4}$$

The physical meaning of each of the one-parameter subgroups in (4.4) can be readily ascertained when we put all others to their identity values. Thus y can be seen to represent coordinate translations $(q \to q + y)$, $-b$ time translations $(t \to t - b)$, $a = d^{-1}$ space—time scale transformations $(q \to q/a,\ t \to t/a^2)$, z phase multiplication $(u \to \exp(iz)u)$, x Galilean transformations $(q \to q + xt,\ u \to \exp(ixq)u)$, c conformal transformations $(q \to q/(1 - ct),\ t \to t/(1 - ct),\ u \to (1 - ct)^{-1/2} \exp\{i[cq^2/(1 - ct)]\}u)$. The last two are not "inspectionally" obvious symmetries of the equation.

If we further require that, under the transformation $\mathcal{J}$, q and t remain real and u remains in $\mathcal{L}^2(R)$, the values of the parameters $a, b, \ldots, z$ must be real. Thus the symmetry group of the free-particle Schrödinger equation is the six-parameter WSL(2, R) group.

D. The results for the heat equation can be read off (4.4) when we replace $t \to 2it$. It is convenient to define $\beta \equiv \frac{1}{2}ib$, $\gamma \equiv -2ic$, $\xi \equiv -2ix$, $\zeta \equiv -iz$. Here we require q, t, and u to be real. In terms of the new variables, we can see that the symmetry group of the heat equation is

TABLE III. Action of the general group transformation $g = \{A, \omega\} \in \mathrm{WSL}(2,C)$ on a function $u(q,t)$, solution of $H^\omega u - i\partial_t u$ for $\omega = h, r, l$ or f, as given by Eq. (4.2).

ω	time transformation	geometrical transformation
h	$\tan\bar{t} = \dfrac{d\tan t - b}{a - c\tan t}$	$\bar{a} = (a\cos t - c\sin t)/\cos\bar{t} = (d\sin t - b\cos t)/\sin\bar{t}$ $\bar{c} = (c\cos t + a\sin t - \bar{a}^{-1}\sin\bar{t})/\cos\bar{t}$ $(\bar{x},\bar{y}) = (x,y)\begin{pmatrix}\cos\bar{t} & \sin\bar{t}\\ -\sin\bar{t} & \cos\bar{t}\end{pmatrix}$, $\bar{Z} = Z$
r	$\tanh\bar{t} = \dfrac{d\tanh t - b}{a - c\tanh t}$	$\bar{a} = (a\cosh t - c\sinh t)/\cosh\bar{t} = (d\sinh t - b\cosh t)/\sinh\bar{t}$ $\bar{c} = (c\cosh t - a\sinh t + \bar{a}^{-1}\sinh\bar{t})/\cosh t$ $(\bar{x},\bar{y}) = (x,y)\begin{pmatrix}\cosh\bar{t} & \sinh\bar{t}\\ \sinh\bar{t} & \cosh\bar{t}\end{pmatrix}$, $\bar{Z} = Z$
l	$\bar{t} = \dfrac{dt - b}{a - ct}$	$\bar{a} = a - ct$, $\bar{c} = c$, $\bar{Z} = Z$ $(\bar{x},\bar{y}) = \{(x,y) + (-t, \tfrac{1}{2}t^2)\begin{pmatrix}a & b\\ c & d\end{pmatrix} + (\bar{t}, \tfrac{1}{2}\bar{t}^2)\}\begin{pmatrix}1 & \bar{t}\\ 0 & 1\end{pmatrix}$
f	$\bar{t} = \dfrac{dt - b}{a - ct}$	$\bar{a} = a - ct$, $\bar{c} = c$, $\bar{Z} = Z$ $(\bar{x},\bar{y}) = (x,y)\begin{pmatrix}1 & \bar{t}\\ 0 & 1\end{pmatrix}$

given by the subgroup of $\mathrm{WSL}(2,C)$ represented by the matrices

$$\left\{\begin{pmatrix} a & -2i\beta \\ \frac{1}{2}i\gamma & d \end{pmatrix},\ (\tfrac{1}{2}i\xi, y, i\zeta)\right\},\quad ad - \beta\gamma = 1 \qquad (4.5a)$$

with $\alpha, \beta, \ldots, \zeta$ real.[32]

The operators which represent the canonical transformation (4.5) in (2.8) will be *bounded* when

$$a \geq 0,\quad \beta \geq 0,\quad \gamma \geq 0,\quad d \geq 0,\quad \xi = 0 \text{ when } \gamma = 0. \qquad (4.5b)$$

The transformations (4.5a) with the restrictions (4.5b) form a *semigroup*, the $\mathrm{SL}(2,R)$ part of which is identical with the $\mathrm{HSL}(2,R)$ semigroup introduced in Ref. 7. It lies on the same orbit—through complex transformations—as the semigroup of real transformations in $\mathrm{SL}(2,R)$ with nonnegative matrix elements.[33] It is here augmented by the Weyl group and can be seen to be a subsemigroup of (4.5a) which preserves the positivity of the time displacements.[34]

E. The treatment of the four quantum Hamiltonians chosen in the last section as orbit representatives, follows the procedure of Eqs. (4.2a,b). We give in Table III the expressions for the time and geometric transformations as done in (4.2). It should be noted, though, that the physical transformations represented by the parameters $a, b, \ldots, y$ differ from case to case.

F. In solving a differential equation, we usually have to contend with *boundary conditions* $u_0(q,t)$ on the boundaries $\beta(q,t) = \text{const}$. Similarity methods choose the transformation $\mathcal{T}_g^{(t)}$ to leave these boundaries invariant: $\beta(q,t) = \beta(\bar{q},\bar{t})$. We will now show that the separating coordinates $(v(q,t),\ t'(t))$ of subsection III.G provide such boundaries in the form $v(q,t) = \text{const}$. Consider an example: Assume the transformation $\mathcal{T}_g^{(t)}$ in (4.3) is of the particular kind $\mathcal{T}_{g(\alpha)}^{(t)} = H_\alpha^{h(t)}$ as in (3.4a). Then (4.3) tells us that $\bar{q} = q/(\cos\alpha - t\sin\alpha)$ and $\bar{t} = (t\cos\alpha + \sin\alpha)/(\cos\alpha - t\sin\alpha)$. Taking the lead from the entries f—h of Tables I and II, we can verify that $v \equiv q(1+t^2)^{-1/2} = \bar{q}(1+\bar{t}^2)^{-1/2} = \text{const}$, while for $t = \tan t'$ and $\bar{t} = \tan\bar{t}'$, $\bar{t}' = t' + \alpha$ simply. Hence, the family of hyperbolae $q^2 = v^2(1+t^2)$ for any $v \in R$ remains invariant under $\mathcal{T}_{g(\alpha)}^{(t)}$.

G. The general proof of this fact hinges on writing $\mathcal{T}_{g(\alpha)}^{(t)} = H_\alpha^{\omega'(t)}$ for some generating operator $H^{\omega'}$. If now we are looking at the solution $u(q,t) = H_t^\omega u(q)$, we should write $H_t^\omega = \mathcal{G}_t H_{t'}^{\omega'}$ and look for the corresponding separation of variables $\{v(q,t), t'(t)\}$ as done in (3.11). The action of $H_\alpha^{\omega(t)}$ will thus be $t' \to t' + \alpha$, and leave $v(q,t)$ as a family of invariant lines on the q—t plane.

H. As for the inverse problem, if we know $\{v(q,t),\ t'(t)\}$ to be system of coordinates where the operator H^ω is separated by a second operator $H^{\omega'}$ [see Eqs. (3.7), (3.11), and (3.12)] with a multiplier $S(v,t)$, then $\bar{v} \equiv v(\bar{q},\bar{t})$, $\bar{t}' \equiv t'(\bar{t})$ as given by (4.2b) will be the separating coordinates of H^ω by the operator $gH^{\omega'}g^{-1}$ with multiplier $S(\bar{v},\bar{t})$. In order to see this, let $\mathcal{T}_g^{(t)}$ [the two-variable representation (4.1)—(4.2) of a transformation g associated with the time development H_t^ω] act on (3.11). The result of this action will still be a solution of (3.2) for H^ω:

$$\begin{aligned}&\mathcal{T}_g^{(t)} H_t^\omega \psi_\lambda^{\omega'}(q)\\ &\quad \equiv H_{\bar{t}}^\omega \mathcal{T}_g \psi_\lambda^{\omega'}(q)\\ &\quad = \exp\{i[S(\bar{v},\bar{t}) - \tfrac{1}{2}\bar{y}(\bar{x} - \bar{c}\bar{a}\bar{y}) + \bar{z} + \lambda\bar{t}']\}\bar{a}^{-1/2}\psi_\lambda^{\omega'}(\bar{v}),\end{aligned} \qquad (4.6)$$

where all barred variables depend on $\bar{q}$ and $\bar{t}$, while $\mathcal{T}_g\psi_\lambda^{\omega'}(q)$ is an eigenfunction of $gH^{\omega'}g^{-1}$.

I. As an illustration, we can apply this relation to the separable solutions of the heat equation seen in subsections III.I and III.J. When the separating operator is H^h (see entries f—h in Table II with $t \to 2it$), then $v = q(1-4t^2)^{-1/2}$. Hence, when we use gH^hg^{-1} to separate, the corresponding variables are

$$\bar{v} = \frac{q + t(\gamma y - d\xi) + (ay - \beta\xi)}{[t^2(\gamma^2 - 4d^2) + 2t(a\gamma - 4d\beta) + (a^2 - 4\beta^2)]^{1/2}},\quad \bar{t} = \frac{dt + \beta}{a + \gamma t}, \qquad (4.7)$$

with $a, \beta, \ldots, y$ as in (4.5). Now, the Hermite separating operator (3.15) is related to H^h through $D = gH^hg^{-1}$ [$g = g_h^{-1}$ as defined below Eq. (3.15a)]; hence, for $g = \{\left(\begin{smallmatrix}1 & 0\\ -i & 1\end{smallmatrix}\right), (0)\}$, $a = 1 = d$, $\gamma = -2$, $\xi = 0 = y$ the separating variables (4.7) become precisely those of (3.15c), namely $\bar{v} = q(1-4t)^{-1/2}$. Conversely, proposing a form for $\bar{v}$, we can find the group element which takes the separating operator to one of the four orbit representatives. We must compare the proposed form with (4.7) and the corresponding expressions for the r, l, and f orbits, solving (nonlinear) algebraic equations for the parameters of g. If these equations are incompatible, the separating operator does not lie on the proposed orbit. If two operators are related through a similarity transformation in the symmetry group of the differential equation of a third, the variables they separate in the latter can be called *equivalent* in a general sense.[35]

Hence, while in Sec. III we found the separating variables for any given operator in the algebra; here we have solved the converse problem.

V. EQUATIONS CONTAINING TERMS IN q^{-2}

A. A class of operators containing terms in q^{-2} is amenable to a treatment parallel to the previous sections. The analysis is in fact simpler, and much of the groundwork has been done in Refs. 2 and 3, so only the general outline and conclusions will be presented. The operators we are refering to are

$$J_1=\frac{1}{4}\left(-\frac{d^2}{dq^2}+\frac{\mu}{q^2}-q^2\right), \tag{5.1a}$$

$$J_2=-\frac{i}{2}\left(q\frac{d}{dq}+\frac{1}{2}\right), \tag{5.1b}$$

$$J_3=\frac{1}{4}\left(-\frac{d^2}{dq^2}+\frac{\mu}{q^2}+q^2\right), \tag{5.1c}$$

which, together with $\mathbb{1}$ close onto an $o(2)\oplus sl(2,R)$ algebra as (2.6), the commuting $o(2)$ is the one generated by $\mathbb{1}$. The operators (5.1) can be seen as the radial part of (2.5) for n-dimensional vectors $\mathbf{Q}$ and $\mathbf{P}$ in the space of angular momentum L, with $\mu=(\frac{1}{2}n+L-1)^2-\frac{1}{4}$ and subjected to a similarity transformation with the factor $|q|^{(n-1)/2}$ in order to cancel the term $[(n-1)/q]d/dq$ in $\mathbf{P}^2$. The operators (5.1) are densely defined and have self-adjoint extensions[3,6] for the ranges of μ specified below, in $\mathcal{L}^2(R^+)$. There is no underlying Weyl algebra here.[2]

Define now k through

$$\mu=(2k-1)^2-\tfrac{1}{4},\quad 2k=1\pm(\mu+\tfrac{1}{4})^{1/2} \tag{5.2}$$

so that the Casimir invariant for the algebra (5.1) can be seen to be $k(1-k)$.

Exponentiating the algebra (5.1) to an $O(2)\otimes SL(2,R)$ group, we associate a realization through 2×2 matrices as in (2.7). As the $O(2)$ part generated by $\mathbb{1}$ corresponds to over-all phase transformations, it is rather trivial and we shall work henceforth with the $SL(2,R)$ part only. The action of the $SL(2,R)$ group on $f\in\mathcal{L}^2(R^+)$ is[2,3,6,7]

$$[C\begin{pmatrix}a&b\\c&d\end{pmatrix}f](q)=b^{-1}\exp(i\pi k)\int_0^\infty dq'(qq')^{1/2}\exp[(i/2b)(aq'^2+dq^2)]$$
$$\times J_{2k-1}(qq'/b)f(q') \tag{5.3a}$$

and, when $b=0$, we have the geometric transformation

$$[C\begin{pmatrix}a&0\\c&a^{-1}\end{pmatrix}f](q)=|a|^{-1/2}\exp[(ica/2|a|^2)q^2]f(|a|^{-1}q), \tag{5.3b}$$

which, save for the absolute values, is identical with (2.8b). The transformations for complex group parameters and the definitions of Hilbert spaces into which these transformations are unitary was detailed in Ref. 2.

B. The adjoint action of $SL(2,R)$ on the algebra is found exactly as in Sec. II. It is represented as in (2.11)–(2.13):

$$K=\sum_j\eta_jJ_j\xrightarrow{g}K'=\sum_j\sum_k\eta_jN_{jk}J_k=\sum_k\eta'_kJ_k$$

where $\|N_{jk}\|$ is the 3×3 upper-left submatrix of (2.12).

The orbit structure of $SL(2,R)$ is well known: there are three orbits corresponding to the sign of the invariant $\Theta=\eta_3^2-\eta_1^2-\eta_2^2$. The orbit representatives are chosen to be $2J_3$ ($\Theta>0$), $2J_1$ ($\Theta<0$), and J_1+J_3 ($\Theta=0$), corresponding respectively to the Schrödinger Hamiltonians for harmonic oscillator plus centrifugal force, repulsive oscillator plus centrigural force, and pure centrifugal force. The relative strength of the oscillator and centrifugal parts can be varied through dilatation transformations in $SL(2,R)$ and the transformations leading a general operator K to one of the orbit representatives are calculated through the use of (2.14b), (2.15b), and (2.16b) excluding the expressions for x, y, θ_4, and θ_5.

For completeness, we list the eigenfunctions and spectrum of the orbit representatives[3]:

Harmonic Oscillator $+\mu/q^2$, $K^h=2J_3$, *spectrum* $\lambda=2(n+k)$, $n=0,1,2,\cdots$:

$$\phi^h_\lambda(q)=[2n!/\Gamma(n+2k)]^{1/2}\exp(-q^2/2)q^{2k-1/2}L_n^{(2k-1)}(q^2). \tag{5.4}$$

Repulsive Oscillator $+\mu/q^2$, $K^r=2J_1$, *spectrum* $\lambda\in R$:

$$\phi^r_\lambda(q)=(2\pi q)^{-1/2}\exp(i\pi k)\exp(\pi\lambda/4)\,2^{i\lambda/2}$$
$$\times[\Gamma(k+\tfrac{1}{2}i\lambda)/\Gamma(2k)]M_{i\lambda/2,k-1/2}(-iq^2), \tag{5.5}$$

where $M_{\mu\nu}$ is the Whittaker function.[24]

Pure Centrifugal, μ/q^2, $K^f=J_1+J_3$, *spectrum* $\frac{1}{2}\lambda^2$, $\lambda\in R^+$:

$$\phi^f_\lambda(q)=(\lambda q)^{1/2}J_{2k-1}(\lambda q). \tag{5.6}$$

These functions are orthogonal and complete for $\mathcal{L}^2(R^+)$. It should be noted that the ϕ^ω_λ are, up to a phase, functions of $|q|$ and in fact $\phi^\omega_\lambda(e^{i\pi}q)=\exp[i\pi(2k-1/2)]\phi^\omega_\lambda(q)$. The operators (5.1) are invariant under $q\to -q$. Thus, the analysis of the eigenfunctions for $q\in R$ and harmonic analysis for functions in $\mathcal{L}^2(R)$ makes use of (5.4)–(5.6) with a few extra facts[3,36]:

(i) For $\mu\geq\frac{3}{4}$ (repulsive centrifugal force), the operators (5.1) have unique self-adjoint extensions in $\mathcal{L}^2(R^+)$ so that $k=\frac{1}{2}(1+[\mu+\frac{1}{4}]^{1/2})\geq 1$ and $\phi^\omega_\lambda(0)=0$.

(ii) For $\frac{3}{4}>\mu>0$ (repulsive), we have two square-integrable solutions and $\phi^\omega_\lambda(q)\sim q^{2k-1/2}$ at $q\to 0$, one for $k_1=\frac{1}{2}(1+[\mu+\frac{1}{4}]^{1/2})$, $\frac{3}{4}<k_1<1$, where the solutions are regular at the origin and one for $k_2=\frac{1}{2}(1-[\mu+\frac{1}{4}]^{1/2})$, $0<k_2<\frac{1}{4}$, where the solutions are irregular, but still square-integrable. We thus have to impose an extra boundary condition at $q=0$. (For example, if we have an infinite potential wall for $q\leq 0$, only the first solutions are acceptable). In $\mathcal{L}^2(R)$, the two families of solutions must be considered.

(iii) At $\mu=0$ the centrifugal "barrier" has disappeared, $k_1=\frac{3}{4}$ and $k_2=\frac{1}{4}$ represent the odd and even solutions, which become zero and constant as $q\to 0$. Their union gives back the spectrum and eigenvalues of the corresponding operators (2.5) on the whole of R.

(iv) For $-\frac{1}{4}<\mu<0$ (attractive centrifugal force), $\frac{1}{2}<k_1<\frac{3}{4}$ and $\frac{1}{4}<k_2<\frac{1}{2}$. Both solutions are regular at the origin. At $\mu=-\frac{1}{4}$ they coalesce.

(v) The centrifugal part cannot be more attractive

than $\mu = -\frac{1}{4}$; otherwise, the k's become $\frac{1}{2} \pm i\nu$ (ν real): the spectrum of K^h is no longer lower-bound and the functions belong to the principal series rather than the lower-bound "discrete" representations of SL(2,R).

From these observations, eigenfunctions of any other operator K in sl(2,C) can be constructed as in II. J as a geometric transform of the eigenfunctions of their orbit representatives.

C. When we come to analyze differential equations of the type

$$Ku(q,t) = -i\frac{\partial}{\partial t}u(q,t) \tag{5.7}$$

with K in the algebra $o(2,C) \oplus sl(2,C)$ generated by (5.1) the time-evolution transforms associated with K can be constructed out of the basis (3.4a, c, d) (the linear potential does not appear here). Copying Sec. II. E we can describe the time evolution of a function, solution of (5.7), expanded in terms of eigenfunctions of an operator K'. In particular the example II. E applies (replacing ψ by ϕ) for $K = 2J_3$ and $K' = J_1 + J_3$ with no change at all. Here we have three instead of the four cases of former sections and Tables I, II, and III on separating coordinates and multipliers apply here when we take out the l-rows and columns. The geometrical action of a is replaced by $|a|$.

Following the results of Sec. IV, we can see that the full invariance group of the class of differential equations (5.7) is the four-parameter group $O(2) \otimes SL(2,R)$ when the appropiate reality and square-integrability conditions are imposed. The illustration in subsection IV. C is valid for the Schrödinger equation with a μ/q^2 potential when we eliminate the variables x and y, and its invariant boundaries are found as in the ensuing discussion.

VI. CONCLUSION

A. First, we would like to compare our approach with that of the "kinematical" invariance groups of Niederer and Boyer. We have dealt with representations of WSL(2,C) on spaces of functions $u(q)$ on the real line q. The time development of a system (3.2) is a particular one-dimensional subgroup of such transformations: $u(q,t) = \mathcal{H}_t u(q)$. Then, we found that the action of WSL(2,C) on the space of functions of two variables could be written as $u(q,t) \xrightarrow{g} v(q,t) = \mathcal{T}_g^{(t)} u(q,t)$ as in (4.1)–(4.2). Clearly $\mathcal{T}_g^{(t)} \equiv \mathcal{H}_t \mathcal{T}_g \mathcal{H}_t^{-1}$. If these transformations are generated as $\mathcal{T}_{g(\alpha)} = \exp(i\alpha F)$ and $\mathcal{T}_{g(\alpha)}^{(t)} = \exp(i\alpha F^{(t)})$, then also $F^{(t)} = \mathcal{H}_t F \mathcal{H}_t^{-1}$, so that F and $F^{(t)}$ are the Schrödinger and Heisenberg pictures of the same operators,[37] while $u(q,t)$ and $u(q)$ are the corresponding wavefunctions. We have

$$[H + i\partial/\partial t,\ F^{(t)}] = 0. \tag{6.1}$$

B. It should be noticed that $F^{(t)}$ generates geometric transformations in q–t space, i. e., $v(q,t)$ is a multiplier function times the function u of the transformed arguments $\bar{q}$ and $\bar{t}$. Thus $F^{(t)}$ can also be realized as a *first-order* differential operator in q and t. Indeed, if now, whenever H appears as a summand in $F^{(t)}$ we replace it by $-i\partial/\partial t$ in such a way that the resulting operator $F^{[t]}$ contain no second-order derivative terms in q and $F^{(t)} - F^{[t]} = f(t)(H + i\partial/\partial t)$, where $f(t)$ is a function only of t which appears among the matrix elements in the representation of $\mathcal{H}_t$ through (2.12). We will have $-[i\partial_t, F^{[t]}] = G^{[t]}$, where $G^{[t]}$ is in the algebra and has similarly H replaced by $-i\partial/\partial t$ and no second-order derivative terms. Now, it is still true that $[H, F^{[t]}] = G^{(t)}$ since H commutes with the H part in $F^{(t)}$. Hence, for some function $g(t)$ which we can find in (2.12),

$$[H + i\partial/\partial t,\ F^{[t]}] = G^{(t)} - G^{[t]} = g(t)(H + i\partial/\partial t), \tag{6.2}$$

acting on the space of differentiable functions of q and t.

Equation (6.2) can be recognized as the starting point for Niederer[9] who proposed definite forms for H (free particle and harmonic oscillator), and Boyer,[10] who left H in the general form $\frac{1}{2}P^2 + V(Q)$ and then determined the possible two-variable first-order differential operators $F^{[t]}$ satisfying (6.2). It was then found that only potentials of the form studied here allowed such a kinematical invariance group.[38,39] A wider class of time-dependent operators, not necessarily polynomials in P and Q have been considered by Anderson, Wulfman, *et al.*[40]

C. Boyer[10] pursued the study of (6.2) for n-dimensional systems and found the symmetry algebra (and group) to be subgroups of $W_n \otimes (SO(n) \otimes SL(2,R))$, called the Schrödinger group. Our method appears applicable to quadratic operators of the type

$$\sum\sum \alpha_{ij} P_i P_j + \sum\sum \beta_{ij}(P_i Q_j + Q_j P_i) + \sum\sum \gamma_{ij} Q_i Q_j + \sum \delta_i Q_i + \sum \epsilon_i P_i + \eta. \tag{6.3}$$

The symmetry algebra will be generated by the operators appearing in the summands and the generated group will be WSp($2n,R$), complexified. This group contains the Schrödinger group but cannot appear out of the starting equation (6.2) since the transformations in WSp($2n,R$) which are not in the Schrödinger group are not geometric transformations in q–t space and hence are not representable as first-order differential operators in these variables satisfying (6.2).

D. Our analysis should reduce the examination of the symmetry group of quadratic Hamiltonians of the type (6.3) to the complete orbit analysis of WSp($2n,R$) or of different real forms of its complex algebra.[41] Presence of "centrifugal force barriers," radial or plane, would cut down the full symmetry and some of the more interesting cases up to three dimensions have been analyzed through separation of variables in the conventional way.[15,16,18] Further, one need not restrict oneself to $\mathcal{L}^2(R^n)$ spaces of functions, but use any differentiable group coset manifold[17] and look for finite- or infinite-dimensional subalgebras in the enveloping algebra[42] of the group. Eventually, one would also like to extend the application of the global group method through matrix algebra (on an extended space, if possible), to other types of differential equations.

ACKNOWLEDGMENTS

I would like to thank the departments of Physics and Astronomy, and of Applied Mathematics of Tel-Aviv University, Israel, where this work was started, for

their hospitality. It is a pleasure to acknowledge illuminating discussions with David Alcaraz, George W. Bluman, and Charles P. Boyer.

[1]K. B. Wolf, J. Math. Phys. 15, 1295 (1974).
[2]K. B. Wolf, J. Math. Phys. 15, 2102 (1974).
[3]C. P. Boyer and K. B. Wolf, J. Math. Phys. 16 (to be published).
[4]R. Gilmore, *Lie Groups, Lie Algebras and Some of Their Applications* (Wiley, New York, 1974).
[5]S. Lie, Arch. Math. (Kristiana) 6, 328 (1881).
[6]M. Moshinsky and C. Quesne, J. Math. Phys. 12, 1772, 1780 (1971); M. Moshinsky, T. H. Seligman, and K. B. Wolf, J. Math. Phys. 13, 901 (1972); M. Moshinsky, SIAM J. Appl. Math. 25, 193 (1973).
[7]P. Kramer, M. Moshinsky and T. H. Seligman, "Complex Extensions of Canonical Transformations in Quantum Mechanics," in *Group Theory and Its Applications, Vol. III*, edited by E. M. Loebl (Academic, New York, 1975).
[8]V. Bargmann, Commun. Pure Appl. Math. 14, 187 (1961); 20, 1 (1967); P. J. Sally, Jr., Mem. Amer. Math. Soc. 69, (1967); C. Itzykson, J. Math. Phys. 10, 1109 (1969); V. Bargmann, "Group Representations in Hilbert Spaces of Analytic Functions" in *Analytical Methods in Mathematical Physics*, edited by R. P. Gilbert and R. G. Newton (Gordon and Breach, New York, 1970); V. Bargmann, P. Butera, L. Girardello, and J. R. Klauder, Rep. Math. Phys. 2, 221 (1971); A. O. Barut and L. Girardello, Commun. Math. Phys. 21, 41 (1971).
[9]U. Niederer, Helv. Phys. Acta 45, 802 (1972); 46, 191 (1973).
[10]C. P. Boyer, Helv. Phys. Acta 47, 589 (1974).
[11]L. V. Ovsjannikov, *Gruppovje Svoistva Diferentsjalny Uravnjeni* (Novosibirsk, Moscow, 1962).
[12]G. W. Bluman and J. D. Cole, J. Math. Mech. 18, 1025 (1969); G. W. Bluman, Quart. Appl. Math. 31, 403 (1974); G. W. Bluman and J. D. Cole, *Similarity Methods for Differential Equations* (Springer, New York, 1974).
[13]P. Winternitz, Ya. A. Smorodinskiĭ, M. Uhlíř, and I. Friš, Sov. J. Nucl. Phys. 4, 444 (1967).
[14]E. G. Kalnins and W. Miller Jr., J. Math. Phys. 15, 499 (1974).
[15]C. P. Boyer, E. G. Kalnins and W. Miller Jr., J. Math. Phys. 16, 499, 512 (1975); C. P. Boyer, Comunicaciones Técnicas CIMAS 5, No. 75 (1974) (to be published in SIAM J. Math. Anal.); C. P. Boyer and K. B. Wolf, Comunicaciones Técnicas CIMAS 6, No. 83 (1975) (to be published in J. Math. Phys.); C. P. Boyer and E. G. Kalnins, work in progress.
[16]E. G. Kalnins, SIAM J. Math. Anal. (April 1975); E. G. Kalnins and W. Miller Jr., J. Math. Phys. 15, 1263 (1974); E. G. Kalnins and W. Miller Jr., Minnesota Univ., School of Mathematics preprint, unnumbered (1975) (to be published in J. Math. Phys.); E. G. Kalnins and W. Miller Jr., Centre de Recherches Mathématiques, Montréal, Preprints CRM-467 and CRM-489 (1975).
[17]E. G. Kalnins, W. Miller Jr. and P. Winternitz, Centre de Recherches Mathématiques, Montréal, Preprint CRM-416 (1974) (to be published in SIAM J. Appl. Math.); E. G. Kalnins, J. Patera, R. T. Sharp, and P. Winternitz, "Two-variable Galilei Group Expansions of Non-relativistic Scattering Amplitudes," in *Group Theory and Its Applications, Vol. III*, edited by E. M. Loebl (Academic, New York, 1975).
[18]This is true in the general sense for the one-dimensional case (compare with Ref. 14). For higher-dimensional spaces some "nonsubgroup" separating coordinates appear (see Refs. 15—17), as in J. Patera and P. Winternitz, J. Math. Phys. 14, 1130 (1973); it is not yet clear whether, when extended to more dimensions, our method will account for them.
[19]K. B. Wolf, "The Heisenberg—Weyl Ring in Quantum Mechanics," in *Group Theory and Its Applications, Vol. III*, edited by E. M. Loebl (Academic, New York, 1975).
[20]P. A. M. Dirac, *The Principles of Quantum Mechanics* (Oxford U. P., London, 1958), 4th ed.
[21]A. Joseph, J. Math. Phys. 13, 351 (1973), Theorem 4.5.
[22]Ref. 1, Sec. 5.
[23]P. A. Mello and M. Moshinsky, Rev. Mex. Fis. 22, 257 (1973).
[24]*Handbook of Mathematical Functions*, edited by M. Abramowitz and I. A. Stegun (Natl. Bureau of Standards, Washington, D. C., 1964).
[25]G. N. Watson, *A Treatise on the Theory of Bessel Functions* (Cambridge U. P., Cambridge, 1966), 2nd ed. Sec. 6.4; Ref. 24, Eq. 10.4.32.
[26]F. Trèves, Bull. Soc. Math. France 97, 193 (1969); M. Miller and S. Steinberg, Commun. Math. Phys. 24, 40 (1971).
[27]The f row in Table II can be compared with the results of Ref. 14, Table I: the columns h, r, l, and f of the former correspond respectively to entries 6 ($d=0$), 5 ($d=0$), 2 ($a=2$), and 1 ($c=0$); in the table of Ref. 14, entries (1,2,4) and (3,5) are equivalent.
[28]Compare with the separable solutions given in the book by Bluman and Cole, Ref. 12, pp. 215—18.
[29]Columns f, l, and r in our Table II correspond to entries 1, 4, and 3 of Ref. 14, Table 2; entry 2 is on the same orbit as column h.
[30]P. Hartman and A. Wintner, Amer. J. Math. 72, 367 (1950); P. C. Rosenbloom and D. V. Widder, Trans. Amer. Math. Soc. 92, 220 (1959); D. V. Widder, Duke Math. J. 29, 497 (1962); G. G. Bilodeau, SIAM J. Math. Anal. 5, 43 (1974).
[31]Ref. 1, Eq. (3.8b).
[32]Our parameters $(\alpha,\beta,\gamma,d;\xi\ y\ \zeta)$ correspond to Bluman and Cole's, $([\beta+1]^{-1},\alpha,-\gamma,\beta+1;-\delta,\kappa,-\lambda)$, in Refs. 12.
[33]Ref. 7, Eqs. (5.32)—(5.34).
[34]The fact that the global transformations constitute a semigroup rather than a group seems not to have been noticed in Refs. 12; this, of course, depends on the space of functions $u(x,t)$ belongs to. Infinitely-differentiable functions of growth $\lesssim \exp(-q^2/4t')$ can be regressed back up to time $t>-t'$. Some of these aspects of canonical transforms are currently under investigation. See also the work of D. V. Widder in Refs. 30.
[35]W. Miller Jr. *et al.* (Refs. 14—16) call two separating coordinate systems *equivalent* if they are related through 'inspectional' transformations (i.e., as translations, dilatations and Galilei transformations, *not* through conformal transformations).
[36]E. C. Titchmarsh, "Eigenfunction Expansions Associated with Second Order Differential Equations" (Oxford Univ. Press, 1962), Vols. I and II; T. Kato, "Perturbation Theory for Linear Operators (Springer, New York, 1966).
[37]M. Moshinsky, S. Kumei, T. Shibuya, and C. E. Wulfman, Instituto de Física, University of México preprint IFUNAM 72-4 (unpublished).
[38]This is also true regarding raising operators, see C. P. Boyer and W. Miller Jr., J. Math. Phys. 15, 1484 (1974).
[39]The generators found in Refs. 12 correspond to ours when we replace their $u(\partial/\partial u)$ by $\mathbb{1}$, as can always be made in the case of linear differential equations.
[40]R. L. Anderson, S. Kumei, and C. E. Wulfman, Phys. Rev. Lett. 28, 988 (1972); R. L. Anderson, S. Kumei and C. E. Wulfman, Rev. Mex, Fís. 21, 1, 35 (1972); J. Math. Phys. 14, 1527 (1973); R. L. Anderson, T. Shibuya and C. E. Wulfman, Rev. Mex. Fís. 23, 257 (1974).
[41]J. G. F. Belinfante and P. Winternitz, J. Math. Phys. 12, 1041 (1971); J. F. Cornwell, Rep. Math. Phys. 2, 239 (1971); 2, 289 (1971); 3, 91 (1972); G. Burdet, M. Perrin and P. Sorba, Commun. Math. Phys. 34, 85 (1973); J. M. Eckins and J. F. Cornwell, Rep. Math. Phys. 5, 17 (1974); J. Patera, P. Winternitz and H. Zassenhaus, J. Math. Phys. 15, 1378, 1932 (1974); Centre de Recherches Mathématiques, Montréal, preprints CRM-470 and CRM-471 (to be published).
[42]P. Chand, C. L. Mehta, N. Mukunda, and E. C. G. Sudarshan, J. Math. Phys. 8, 2048 (1967).

The relationship between the normalization coefficient and dispersion function for the multigroup transport equation*

Mitchell J. Feigenbaum†

Department of Physics, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061
(Received 25 July 1974; revised manuscript received 7 November 1975)

An explicit formula for the discrete Case normalization coefficient is presented in terms of functions related to the dispersion function. These functions are easily determined and provide the normalization coefficient without need of prior evaluation of the eigenvectors.

The single-group (reduced) transport equation[1]

$$(z-\mu)\phi(\mu)=\frac{cz}{2}\int_{-1}^{1}d\mu'\,\phi(\mu') \tag{1}$$

is an eigenvalue equation for z: setting

$$\int_{-1}^{1}d\mu\,\phi(\mu)=1, \tag{2}$$

one obtains ($z\notin[-1,1]$)

$$\phi(\mu)=\frac{cz/2}{z-\mu} \tag{3}$$

with z determined by reimposing (2) upon (3):

$$1=\int_{-1}^{1}\phi(\mu)\,d\mu=\frac{cz}{2}\int_{-1}^{1}\frac{d\mu}{z-\mu}\equiv M(z).$$

That is, the eigenvalues z are the zeroes of the function $\Omega(z)$, where

$$\Omega(z)=1-\frac{cz}{2}\int\frac{d\mu}{z-\mu}=1-M(z). \tag{4}$$

Had ϕ been multicomponented (as for multigroup equations), the eigenvalues z would similarly have been the zeroes of a function $\Omega(z)$, which then is the determinant of the coefficients of the linear system analogous to (1):

$$\Omega(z)\equiv\det(\mathbf{I}-\mathbf{M}(z)). \tag{5}$$

That is, (4) is simply the one-dimensional case of (5). We shall write the explicit form of the matrix $\mathbf{M}(z)$ later.

A normalization factor N is defined for a solution to (1), according to

$$N\equiv\int_{-1}^{1}d\mu\,\mu\phi^2(\mu). \tag{6}$$

By utilizing the solution (3) [with z replaced by z_0, where $\Omega(z_0)=0$], it is easy to verify that the value of N satisfies the well-known formula

$$N=\tfrac{1}{2}cz_0^2\Omega'(z_0)=-\tfrac{1}{2}cz_0^2M'(z_0). \tag{7}$$

For multi-group equations for particular (and small, e.g., two) numbers of groups results similar to (7) have surfaced in the literature.[2] In this paper we attempt to determine just what the connection between N and Ω' is for a fairly general class of nonconstant, nonisotropic multigroup equations.

To be exact, we investigate the equation[3]

$$(\Sigma z-\mu\mathbf{I})\cdot\boldsymbol{\phi}(\mu)=z\sum_{m=1}^{\alpha}\mathbf{A}^m(\mu)\cdot\int_{-1}^{1}d\mu'\,\mathbf{B}^m(\mu')\circ\boldsymbol{\phi}(\mu') \tag{8}$$

for an α-fold degenerate nonconstant, nonisotropic scattering kernel, with Σ the diagonal matrix of cross-sections:

$$\Sigma=\begin{pmatrix}\sigma_1 & & & 0\\ & \sigma_2 & & \\ & & \ddots & \\ 0 & & & \sigma_N\end{pmatrix}$$

for an N-group problem.

With $\mathbf{M}(z)$ the $N\alpha\times N\alpha$ matrix,

$$M_{i}^{(m)(n)}{}_{j}(z)\equiv z\int_{-1}^{1}d\mu\,B_{ir}^{(m)}(\mu)(\Sigma z-\mu\mathbf{I})_{rs}^{-1}A_{sj}^{(n)}(\mu), \tag{9}$$

we shall establish that

$$N_i=-z_i^2\lambda_i'(z_i)=z_i^2\Omega'(z_i)\prod_{m\neq i}[1-\lambda_m(z_i)]^{-1} \tag{10}$$

where $\lambda_i(z)$ is ith eigenvalue of $\mathbf{M}$, and

$$\Omega=\det(\mathbf{I}-\mathbf{M})=\prod_{i=1}^{N\alpha}[1-\lambda_i(z)]$$

In analogy to the solution of (1), one solves (8) by isolating $\boldsymbol{\phi}$. Defining

$$\int_{-1}^{1}d\mu\,\mathbf{B}^{(m)}(\mu)\cdot\boldsymbol{\phi}(\mu)\equiv\boldsymbol{\beta}^{(m)} \tag{11}$$

upon multiplying by $(\Sigma z-\mu\mathbf{I})^{-1}$,

$$\boldsymbol{\phi}(\mu)=z(\Sigma z-\mu'\mathbf{I})^{-1}\cdot\sum_{m=1}^{\alpha}\mathbf{A}^{(m)}(\mu)\cdot\boldsymbol{\beta}^{(m)}.$$

Next, multiply by $\mathbf{B}^{(n)}(\mu)$ and integrate over μ:

$$\int_{-1}^{1}\mathbf{B}^{(n)}(\mu)\circ\boldsymbol{\phi}(\mu)\,d\mu=\boldsymbol{\beta}^{(n)}$$
$$=\sum_{m=1}^{\alpha}\left(\int_{-1}^{1}d\mu\,\mathbf{B}^{(n)}(\mu)\cdot z(\Sigma z-\mu'\mathbf{I})^{-1}\cdot\mathbf{A}^{(m)}(\mu)\right)\cdot\boldsymbol{\beta}^{(m)}$$
$$\equiv\sum_{m=1}^{\alpha}\mathbf{M}^{(n)(m)}(z)\circ\boldsymbol{\beta}^{(m)}.$$

That is, $\sum_m(\delta^{nm}\mathbf{I}-\mathbf{M}^{(n)(m)}(z))\boldsymbol{\beta}^{(m)}=0$, where $\mathbf{M}^{(n)(m)}$ is defined by (9). Clearly, this is a usual homogeneous system of equations in $N\alpha$ dimensions. Thus, apart from direct product subscripting, an α-fold degenerate kernel presents the identical mathematical problems as the onefold kernel $\mathbf{A}(\mu)\cdot\mathbf{B}(\mu')$. Accordingly, with no loss of generality, we consider the notationally simpler problem of the onefold degenerate kernel:

$$(\mathbf{I}-\mathbf{M}(z))\circ\boldsymbol{\beta}=0,\quad \boldsymbol{\phi}=z(\Sigma z-\mu\mathbf{I})^{-1}\circ\mathbf{A}\circ\boldsymbol{\beta}, \tag{12}$$

where

$$\mathbf{M}(z)=\int_{-1}^{1}d\mu\,\mathbf{B}(\mu)\cdot z(\Sigma z-\mu\mathbf{I})^{-1}\cdot\mathbf{A}(\mu), \tag{13}$$

$$\boldsymbol{\beta}\equiv\int_{-1}^{1}\mathbf{B}(\mu)\cdot\boldsymbol{\phi}(\mu)\,d\mu. \tag{14}$$

Next, define the adjoint solution $\boldsymbol{\phi}^*$:

$$\boldsymbol{\phi}^*\cdot(\Sigma z-\mu\mathbf{I})=z\Big(\int d\mu'\,\boldsymbol{\phi}^*(\mu')\circ\mathbf{A}(\mu')\Big)\cdot\mathbf{B}(\mu). \tag{15}$$

In an identical fashion to the above, with

$$\boldsymbol{\beta}^*\equiv\int_{-1}^{1}d\mu\,\boldsymbol{\phi}^*(\mu)\cdot\mathbf{A}(\mu) \tag{16}$$

one obtains

$$\boldsymbol{\beta}^*\cdot(\mathbf{I}-\mathbf{M}(z))=0 \quad \text{and}$$

and

$$\boldsymbol{\phi}^*=\boldsymbol{\beta}^*\cdot\mathbf{B}\cdot z(\Sigma z-\mu\mathbf{I})^{-1}. \tag{17}$$

($\boldsymbol{\beta}$ differs from $\boldsymbol{\beta}^*$ only when $\mathbf{M}$ is *nonsymmetric.*) The solubility of either (12) or (17) is exactly the eigenvalue condition $\Omega(z)=0$, where

$$\Omega(z)\equiv\det(\mathbf{I}-\mathbf{M}(z)), \tag{18}$$

with $\boldsymbol{\beta}$ and $\boldsymbol{\beta}^*$, respectively, right and left eigenvectors of $\mathbf{M}$ corresponding to the eigenvalue $+1$; the condition on a z_0 is that $\mathbf{M}(z_0)$ should possess the eigenvalue $+1$.

Corresponding to the mth zero of $\Omega(z)$ [i.e., $\Omega(z_m)=0$] is a $\boldsymbol{\beta}^{(m)}$ and $\boldsymbol{\beta}^{*(m)}$. As a natural normalization for that solution, we choose

$$\boldsymbol{\beta}^{*(m)}\cdot\boldsymbol{\beta}^{(m)}=1 \tag{19}$$

and shortly comment on when this condition is tenable: At this point we cannot yet even comment on orthogonality of different modes. Normalization on the solution through (19), having been set, the normalization coefficient is determined:

$$N_m\equiv\int_{-1}^{1}d\mu\,\mu\boldsymbol{\phi}^*(\mu)\cdot\boldsymbol{\phi}(\mu)$$
$$=z_m{}^2\int_{-1}^{1}d\mu\,\boldsymbol{\beta}^{*(m)}\cdot\mathbf{B}(\mu)\cdot\mu(\Sigma z-\mu\mathbf{I})^{-2}\cdot\mathbf{A}(\mu)\cdot\boldsymbol{\beta}^{(m)}$$
[by (12) and (17)]

and

$$N_m=-z_m{}^2\boldsymbol{\beta}^{*(m)}\circ\mathbf{M}'(z_m)\circ\boldsymbol{\beta}^{(m)} \quad [\text{where } \mathbf{M}'\equiv(d/dz)\mathbf{M}]. \tag{20}$$

Equation (20) establishes some connection between N and $\mathbf{M}$, although it requires the evaluation of both $\boldsymbol{\beta}$ and $\boldsymbol{\beta}^*$ prior to calculating N. It is our goal to provide an evaluation of N independent of explicit $\boldsymbol{\beta}$ dependence. Unfortunately, Eqs. (12) and (17) are not valid for all z: Rather, they are a compatable system of equations only for certain specific values of z (i.e., the z_m). Accordingly, neither of (12) or (17) can be differentiated to be useful in (20). Thus, we are forced to pose a more flexible eigenvalue problem for $\mathbf{M}$. Clearly, for any z, we can evaluate the elements of $\mathbf{M}(z)$ and pose its eigenvalue problem. Equation (12) poses a more restricted problem, in that it seeks out those special values of z for which the eigenvalue is $+1$: For other values of z, $\mathbf{M}$ will possess eigenvalues different from 1 and z-dependent:

$$\mathbf{M}(z)\cdot\boldsymbol{\gamma}(z)=\lambda(z)\boldsymbol{\gamma}(z), \tag{21}$$

where

$$\lambda(z_m)=1.$$

For a given z, there will, in general, be N different eigenvalues:

$$\lambda_m(z), \quad m=1,\ldots,N$$

and, in general, at a z_m satisfying $\Omega(z_m)=0$, only one λ will achieve the value $+1$. Accordingly, we label the z-dependent eigenvalues with the same index that labels the z's that satisfy $\Omega(z)=0$:

$$\lambda_m(z_m)\equiv 1. \tag{22}$$

[Should $\Omega(z)=0$ possess a degenerate root, evidently exactly that number of the λ's must simultaneously achieve the value $+1$ at that z-value.] For z_m, (21) becomes

$$\mathbf{M}(z_m)\cdot\boldsymbol{\gamma}_m(z_m)=\boldsymbol{\gamma}_m(z_m),$$

where $\boldsymbol{\gamma}^{(m)}(z)$ is the eigenvector associated with λ_m. That is,

$$\boldsymbol{\beta}^{(m)}=\boldsymbol{\gamma}_m(z_m). \tag{23}$$

Similarly,

$$\boldsymbol{\gamma}_m^*(z)\cdot\mathbf{M}(z)=\lambda_m(z)\boldsymbol{\gamma}_m^*(z) \tag{24}$$

and

$$\boldsymbol{\beta}^{*(m)}=\boldsymbol{\gamma}_m^*(z_m). \tag{25}$$

We are now in a position to examine orthonormality questions.

$$\mathbf{M}(z)\cdot\boldsymbol{\gamma}_m(z)=\lambda_m(z)\boldsymbol{\gamma}_m(z)$$
$$\Rightarrow\boldsymbol{\gamma}_n^*(z)\cdot\mathbf{M}(z)\cdot\boldsymbol{\gamma}_m(z)=\lambda_m(z)\boldsymbol{\gamma}_n^*(z)\cdot\boldsymbol{\gamma}_m(z)$$

and

$$\boldsymbol{\gamma}_n^*(z)\cdot\mathbf{M}(z)=\lambda_n(z)\boldsymbol{\gamma}_n^*(z)$$
$$\Rightarrow\boldsymbol{\gamma}_n^*(z)\cdot\mathbf{M}(z)\cdot\boldsymbol{\gamma}_m(z)=\lambda_n(z)\boldsymbol{\gamma}_n^*(z)\cdot\boldsymbol{\gamma}_m(z),$$

i.e.,

$$(\lambda_m(z)-\lambda_n(z))\boldsymbol{\gamma}_n^*(z)\cdot\boldsymbol{\gamma}_m(z)=0 \tag{26}$$

so that

$$\boldsymbol{\gamma}_n^*(z)\cdot\boldsymbol{\gamma}_m(z)=0 \quad \text{for } \lambda_m(z)\neq\lambda_n(z). \tag{27}$$

Should all the eigenvalues be distinct, then these $\boldsymbol{\gamma}_m$'s must span the N-dimensional space. Since, by (27), $\boldsymbol{\gamma}_n^*$ is orthogonal to $N-1$ linearly independent vectors, and is nonnull, it must have a projection upon the last, so that by appropriate normalization coefficients of the $\boldsymbol{\gamma}$'s, one can set

$$\boldsymbol{\gamma}_n^*\cdot\boldsymbol{\gamma}_m=\delta_{nm}. \tag{28}$$

Accordingly, by defining

$$G_{im}\equiv(\boldsymbol{\gamma}_m)_i, \quad (\boldsymbol{\gamma}_n^*)_i=G_{ni}^{-1}, \tag{29}$$

where (28) also guarantees G's invertibility. Clearly,

G accomplishes $\mathbf{M}$'s diagonalization

$$\gamma_n^* \cdot \mathbf{M} \cdot \gamma_m = \lambda_m \gamma_n^* \cdot \gamma_m$$
$$\Leftrightarrow \quad (\mathbf{G}^{-1} \cdot \mathbf{M} \cdot \mathbf{G})_{mn} = \lambda_m \delta_{mn} \equiv (\Lambda)_{mn}. \tag{30}$$

However, with degenerate eigenvalues and $\mathbf{M}$ nonsymmetric, diagonalization is not in general possible. Should it be possible, $\mathbf{M}$'s spectrum is termed complete. We assume completeness from this point onwards. This is important because it guarantees the validity of the normalization posited in (19): Set $m=n$ in (28) and evaluate at $z=z_m$:

$$1 = \gamma_m^*(z_m) \cdot \gamma_m(z_m) = \beta^{*(m)} \cdot \beta^{(m)}.$$

Also,

$$\begin{aligned}\Omega(z) &= \det(\mathbf{I} - \mathbf{M}(z)) = \det \mathbf{G}^{-1}(z)\mathbf{G}(z) \cdot \det(\mathbf{I} - \mathbf{M}(z)) \\ &= \det(\mathbf{G}^{-1}(z) \cdot (\mathbf{I} - \mathbf{M}(z)) \cdot \mathbf{G}(z)) \\ &= \det(\mathbf{I} - \Lambda(z)) \\ &= \prod_{m=1}^{N} [1 - \lambda_m(z)].\end{aligned} \tag{31}$$

Since (21) holds for all z, we can differentiate it:

$$\mathbf{M}'(z) \cdot \gamma_m(z) + \mathbf{M}(z) \cdot \gamma_m'(z) = \lambda_m'(z)\gamma_m(z) + \lambda_m(z)\gamma_m'(z)$$

or

$$\mathbf{M}'(z) \cdot \gamma_m(z) = \lambda_m'(z)\gamma_m(z) + (\lambda_m(z)\mathbf{I} - \mathbf{M}(z)) \cdot \gamma_m'(z).$$

Projecting upon γ_m^*, paying attention to (24) and (28), we obtain

$$\begin{aligned}&\gamma_m^*(z) \cdot \mathbf{M}'(z) \cdot \gamma_m(z) \\ &\quad = \lambda_m'(z)\gamma_m^*(z) \cdot \gamma_m(z) + \gamma_m^*(z) \cdot (\lambda_m(z)\mathbf{I} - \mathbf{M}(z)) \cdot \gamma_m'(z) \\ &\quad = \lambda_m'(z).\end{aligned}$$

Finally, evaluating at $z=z_m$,

$$\begin{aligned}&\gamma_m^*(z_m) \cdot \mathbf{M}'(z_m) \cdot \gamma_m(z_m) \\ &\quad = \beta^{*(m)} \cdot \mathbf{M}'(z_m) \cdot \beta^{(m)} = \lambda_m'(z_m)\end{aligned}$$

so that

$$N_m = z_m^2 \lambda_m'(z_m) \quad \text{where } \lambda_m(z_m) = 1. \tag{32}$$

Thus, knowledge of the $\lambda(z)$*'s suffices to determine at once the* z_m*'s and* N_m*'s.* To rewrite (32) in terms of Ω, differentiate (31):

$$\Omega'(z_m) = -\lambda_m'(z_m) \prod_{i \neq m} [1 - \lambda_i(z_m)]$$

or

$$\Omega'(z_m) = (N_m/z_m^2) \prod_{i \neq m} [1 - \lambda_i(z_m)]. \tag{33}$$

It is, at this point, perhaps useful to explicate these ideas by examining a two-group equation with constant, isotropic kernel,[4]

$$\Sigma = \begin{pmatrix} 1 & 0 \\ 0 & \sigma \end{pmatrix}, \quad \mathbf{A} \cdot \mathbf{B} = \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix} \equiv \mathbf{C},$$

$$(\Sigma z - \mu \mathbf{I}) \cdot \phi(\mu) = z\mathbf{C} \cdot \int_{-1}^{1} d\mu\, \phi(\mu'),$$

and

$$\phi^*(\mu)(\Sigma z - \mu \mathbf{I}) = z \int_{-1}^{1} \phi^*(\mu')\, d\mu' \cdot \mathbf{C}.$$

Defining

$$\begin{aligned}\beta &= \int d\mu'\, \phi(\mu'), \\ \beta^* &= \int d\mu'\, \phi^*(\mu') \cdot \mathbf{C},\end{aligned} \tag{34}$$

we have

$$\mathbf{M}(z) = z \int_{-1}^{1} d\mu\, (\Sigma z - \mu \mathbf{I})^{-1} \cdot \mathbf{C} \tag{35}$$

with β and β^* right and left eigenvectors. Writing out (35), we have

$$\begin{aligned}\mathbf{M}(z) &= \begin{pmatrix} z \int_{-1}^{1} d\mu/(z-\mu) & 0 \\ 0 & z \int_{-1}^{1} d\mu/(\sigma z - \mu) \end{pmatrix} \cdot \mathbf{C} \\ &\equiv \begin{pmatrix} f(z) & 0 \\ 0 & (1/\sigma) f(\sigma z)) \end{pmatrix} \cdot \mathbf{C}.\end{aligned} \tag{36}$$

Calculating $\lambda(z)$:

$$\det\left(\lambda \mathbf{I} - \begin{pmatrix} f(z) & 0 \\ 0 & (1/\sigma) f(\sigma z) \end{pmatrix} \cdot \mathbf{C}\right) = 0, \tag{37}$$

which, after some algebra, reduces to

$$\lambda^2 - \lambda[C_{11} f(z) + (C_{22}/\sigma) f(\sigma z)] + (C/\sigma) f(z) f(\sigma z) = 0,$$
$$C \equiv \det \mathbf{C}. \tag{38}$$

Differentiating (38),

$$2\lambda\lambda' - \lambda'[C_{11} f(z) + (C_{22}/\sigma) f(\sigma z)] - \lambda[C_{11} f'(z) + C_{22} f'(\sigma z)] + (C/\sigma) f'(z) f(\sigma z) + C f(z) f'(\sigma z) = 0.$$

Solving for λ' and setting $\lambda = 1$, $z = z_0$,

$$\lambda' = -\frac{(C/\sigma) f'(z_0) f(\sigma z_0) + C f(z_0) f'(\sigma z_0) - [C_{11} f'(z_0) + C_{22} f'(\sigma z_0)]}{2 - C_{11} f(z_0) - (C_{22}/\sigma) f(\sigma z_0)}$$

so that

$$N_0 = -z_0^2 \lambda'(z_0) = C z_0^2 \frac{(1/\sigma) f'(z_0) f(\sigma z_0) + f(z_0) f'(\sigma z_0) - (1/C)[C_{11} f'(z_0) + C_{22} f'(\sigma z_0)]}{2 - C_{11} f(z_0) - (C_{22}/\sigma) f(\sigma z_0)}. \tag{39}$$

To evaluate z_0, one sets $\lambda = 1$ in (38), which of course, is simply $\Omega(z_0) = 0$ as can be seen by setting $\lambda = 1$ in (37). Since f is a perfectly definite function

$$f(z) = z \int_{-1}^{1} d\mu/(z-\mu) = z \ln\left|(1+z)/(1-z)\right| = 2z \tanh^{-1}(1/z),$$

once z_0 is evaluated, N_0 is obtained from (39) without further computation. It is to be recalled here that N_0 of (39) is the normalization factor for the solution normalized to $\beta^* \cdot \beta = 1$, or

$$\left(\int d\mu\, \phi^*(\mu)\right) \cdot \mathbf{C} \cdot \left(\int d\mu\, \phi(\mu)\right) = 1.$$

ACKNOWLEDGMENT

The author is indebted to P. Zweifel for several stimulating remarks, as well as for his suggestion of the problem herein considered.

*Research performed under the auspices of the U.S.E.R.D.A.
†Present address: Los Alamos Scientific Laboratory, Theoretical Division, Mail Stop 210, Los Alamos, New Mexico 87544.

[1]K.M. Case, Ann. Phys. (N.Y.) **9**, 1 (1960).
[2]A comprehensive review has been given by N.J. McCormick and I. Kuscer, Advan. Nucl. Sci. Tech. (N.Y.) **7**, 181 (1973). An extensive list of references is incorporated.
[3]Ref. 2, esp. Ref. 76 in Ref. 2.
[4]C.E. Siewert and P.S. Shieh, J. Nucl. Energy **21**, 385 (1967).

Models of Zermelo Frankel set theory as carriers for the mathematics of physics. I*

Paul A. Benioff

Chemistry Division, Argonne National Laboratory, Argonne, Illinois 60439
(Received 18 February 1975; resubmitted 3 July 1975)

This paper is a first attempt to explore the relationship between physics and mathematics "in the large." In particular, the use of different Zermelo Frankel model universes of sets (ZFC models) as carriers for the mathematics of quantum mechanics is discussed. It is proved that given a standard transitive ZFC model M, if, inside M, $B(H_M)$ is the algebra of all bounded linear operators over a Hilbert space H_M, there exists, outside M, a Hilbert space H and an algebra $B(H)$, along with isometric monomorphisms U_M and V_M from H_M into H and from $B(H_M)$ into $B(H)$. U_M and V_M are used to relate quantum mechanics based on M to quantum mechanics based on the usual ZFC model. It is then shown that, contrary to what one would expect, all ZFC models may not be equivalent as carriers for the mathematics of physics. In particular, it is proved that if one requires that an outcome sequence, associated with an infinite repetition of measuring a question observable on a system prepared in some state, be random, and if a strong definition of randomness is used, then the minimal standard ZFC model cannot be a carrier for the mathematics of quantum mechanics.

I. INTRODUCTION

The basic position taken here and in the succeeding paper is that the relationship between physics and mathematics is deeper and more complex than has perhaps been realized. As a step towards understanding this relationship we take seriously here the fact that all the mathematics used by physics so far, and in fact most of mathematics itself, is considered to take place in intuitive set theory. This can be seen, for example, by examining most any comprehensive treatise in mathematics or mathematical physics where the first few pages usually give a brief review of set theory. Further on in such treatises one also sees such statements as "a Hilbert space is a set such that $\cdots$," the set of complex numbers is . . . ," etc.

There are many ways to treat axiomatically the intuitive concept of set. By far the most extensively studied and developed is Zermelo Frankel set theory or ZF set theory. This theory, which axiomatizes much of the intuitive concept of set, is adequate to encompass most of the mathematics done to date and all of the mathematics used so far by physics. More precisely this means that all the mathematical theorems and results used so far by physics can be cast as theorems and results of ZF set theory. For example the Hilbert space axioms give various properties of vector addition "$+$", scalar multiplication "$\circ$", and the scalar product $(\ ,\ )$. "$+$" is a map from $H \times H$ to H and as such is a set of ordered triples of elements of H. "$\circ$" is a map from $C \times H$ to H and as such is a subset of $C \times H \times H$, and so on.

One often considers in physics many different Hilbert spaces, many different algebras, groups, etc. Formally this is expressed by saying that the axiomatic theory of Hilbert spaces has many models where a model is a mathematical structure in which the axioms are true. Similar statements hold for the theory of algebras, groups, etc.

Now the same also holds true for ZF set theory. That is, there are many collections of sets along with a binary membership relation which satisfy the axioms of set theory. The existence of many different such ZF models or "universes of sets" as they are referred to is, perhaps, unfamiliar. However, it is an essential aspect of modern axiomatic set theory where much of the the work is directed towards the construction of models.

Probably the main reason for this unfamiliarity is that the mathematics used by physics and most of mathematics itself is considered to take place in one universe, V, of sets. [V is that part of the Cantorian universe of sets which is axiomatizable in ZF set theory.] This is clear from the way mathematics is used in physics. An example of this, which will be discussed more later on in Sec. III, is the uniqueness (up to isomorphism) of the scalar field of (standard) complex numbers, which is taken for granted in physics and most mathematics. In this case one carries out the mathematics of physics entirely within V and ignores entirely the existence of the other universes of sets.

The concern of this paper begins with the following point. The different universes of sets, when viewed from the inside have all the same "common" properties. They differ only by such "esoteric" properties as whether or not the continuum hypothesis holds,[1] whether or not the universe is Gödel constructible, etc.

Now all the mathematical theorems and results used by physics so far are independent of these properties. For example, no theorem used in physics so far requires for its proof the continuum hypothesis or its negation, or the axiom of constructibility or its negation, etc. Thus as far as the mathematics of physics is concerned, all universes of sets are entirely equivalent and it should not matter which universe one takes as the mathematical carrier for physics.

In this paper this point is examined in more detail. Among other things, it is suggested that there is a possibility that this may not be true. In Sec. III some aspects of quantum mechanics based in a model of ZF set theory are compared with quantum mechanics based on the usual informal mathematics. Part A gives some

necessary conditions which must be satisfied if a ZF model **M** is to be a carrier for the mathematics of quantum mechanics. In essence these conditions are the strengthened interpretative rules for quantum mechanics given elsewhere.[2] Sections B and C give the necessary correspondences between some of the mathematics of quantum mechanics inside **M** and outside **M**. In particular it is shown in Part B (Theorem 1) that given a set H_M which, inside **M**, is a Hilbert space, there exists a Hilbert space H outside **M** and a natural map U_M to H which is an isometric monomorphism. However, U_M is not in general an isomorphism. Section C gives similar results for the algebras of bounded linear operators over H_M and H with a corresponding map V_M.

In Part D, V_M is used to give a natural correspondence between quantum mechanics based on **M** and quantum mechanics based on the usual informal mathematics.

In Sec. IV the equivalence of all universes of sets as carriers for the mathematics of quantum mechanics is discussed. It is suggested that all models may not be equivalent. In particular it is shown (Theorem 11) that if the definition of randomness used elsewhere[2] is correct, then the minimal standard model of ZF set theory cannot be a carrier for the mathematics of physics.

This result is discussed in Sec. V with respect to different definitions of randomness. In particular it is noted that an important open question is to show how strong the weakest possible definition of randomness must be to not run into any inconsistencies or difficulties. Section II gives a preliminary discussion of those parts of standard mathematical logic and ZF set theory which are relevant to this paper. The discussion on absoluteness in set theory is especially relevant.

II. PRELIMINARIES

A. General

In mathematical logic a theory such as the theory of groups, C^*-algebras, ZF set theory, etc., consists of a set of formulas as part of the language of the theory along with a particular designated subset of the formulas as the nonlogical axioms of the theory.

The symbols of the language consist of the logical symbols $\vee$ (or), $\wedge$ (and), $\neg$ (not), $\exists$ (there exists), (for all) as well as $=$, and some nonlogical relation and function symbols, constant symbols, and a countable set of variables x, y, $\cdots$. The atomic formulas of a language are those symbol strings built up directly from the nonlogical symbols without using any of the logical symbols except $=$. Thus in ZF set theory $u=v$ and $u\in v$ with u, v variables are the atomic formulas. The formulas of a language are defined as the smallest class of symbol strings which contains the atomic formulas and is closed under the following: If ψ and θ are formulas and u is a variable, then $\neg\psi$, $\psi\vee\sigma$, $\psi\wedge\theta$, $\exists u\psi$, and $u\psi$ are formulas.

Note in this definition ψ, θ, u, and v are syntactic variables and are part of the informal language (English) which is used to talk about formulas and variables. They do not belong to the object language.

In a formula a variable is bound if it is acted on by a quantifier; otherwise it is free. Thus, in $\exists x x\in y$, x is bound and y is free. A sentence is a formula with no free variables.

A theory is a language as specified above with a particular set of sentences designated as the nonlogical axioms. Besides the nonlogical axioms one has the logical axioms and deduction rules which are common to all theories. The set of theorems is the smallest set of formulas which contains all the axioms and is closed under the deduction rules.

The above describes the formal grammar rules for a theory. A theory acquires meaning by interpreting it in various mathematical structures. A structure for a language is a universe of objects along with a specific relation and function for each relation and function symbol in the language and a particular element of the universe for each constant symbol. The variables (in a first order theory) range over the entire universe. A model for a theory is a structure for the language of a theory in which each of the axioms is true.

Some important theorems about the relationships between theories and their models are the following: A theory is consistent if and only if it has a model where a theory is defined to be consistent if not all sentences are theorems. A sentence is a theorem if and only if it is true in every model. For almost all theories there is no formula of the theory which formalizes the notion of truth in the models. One always has to go to a stronger theory.

B. ZF set theory

ZF set theory has been greatly developed in the past few years and several texts are available.[1,3-7] Here a few aspects which are most relevant will be reviewed.

The intuitive idea of set is the following.[3] One starts with an empty universe (i.e., the set theoretic vacuum) and forms sets at various stages by iterating a collection process. At each stage one forms all collections of sets formed by the collection process at all earlier stages. One can imagine having an infinite succession of stages. Then there is a following stage consisting of all collections of sets formed at all the stages in the succession. Thus the empty set, 0, is formed at the first stage. The second stage gives the sets $\{0\}$ and 0 and so on. A universe of sets consists of all the sets formed at all stages by iteration of this collection process.[9]

Briefly one version of the ZF axioms is as follows:[1] (1) the axiom of extensionality, which says that two sets which contain the same elements are equal; (2) the existence axiom of the empty sets; (3) the axiom of unordered pairs, which says, given sets x and y, $\{x,y\}$ is a set; (4) the union axiom, which says, given a set x there is a set containing all and only those elements which are members of some element of x; (5) the power set axiom, which says that given any set x there is a set whose elements are the subsets of x; (6) the axiom of infinity, which says that there is a set y which contains 0 and if $x\in y$ so is $x\cup\{x\}\in y$; (7) the axiom of foundation, which says that each set x has a minimal element y in the sense that x and y have no elements in common; (8) the replacement schema, which says that for each

formula $q(x,\underset{\rightarrow}{y},\underset{\rightarrow}{p})$ if for each x there is a unique y for which $q(x,y,\underset{\rightarrow}{p})$ holds, then for each set z there is a set $w=\{y\,|\,x\in z$ and $q(x,y,\underset{\rightarrow}{p})$, where $\underset{\rightarrow}{p}$ stands for a set of parameters; and finally (9) the axiom of choice, which says that for any set x there is a set y whose elements are obtained by choosing exactly one element from each member of x.

The replacement axiom schema consists of an infinite number of axioms, one for each formula, so that ZF set theory has an infinite number of axioms. The axiom of foundation excludes any infinite descending $\in$ chains of the form $\cdots x_{n+1}\in x_n\cdots\in x_1\in x_0$.

In set theory one must distinguish between sets and proper classes. The general object is a class. A set is any class which is an element of another class. A proper class is a class which is not an element of any other class. From the intuitive construction, a set is any class which is generated at some stage of the collection process. If a class is proper, then for each stage of the process there is a later stage which generates some members of the class. The universe of all sets is a proper class. Each formula $q(x)$ of ZF set theory generates a class, by $\{x\,|\,q(x)\}$, which may or may not be proper.

A class is *transitive* if, for all Y, $Y\in X\rightarrow Y\subseteq X$ [capital letters are often used to denote the general object]. An *ordinal* is a transitive set each of whose elements is transitive. The ordinals are well ordered by the membership relation; thus $\alpha<\beta\leftrightarrow\alpha\in\beta$ for any ordinals α,β. Each ordinal $\alpha=\{\beta\,|\,\beta$ is an ordinal and $\beta<\alpha\}$.

In the intuitive construction process a new ordinal is generated at each stage. If α is generated at some stage, the ordinal $\alpha\cup\{\alpha\}=\alpha+1$ is generated at the next stage. The ordinal $\alpha+1$ is the successor ordinal of α. A limit ordinal is an ordinal which is not equal to 0 and which is not the successor of any other ordinal. The natural numbers 0 (the empty set), $1=\{0\}$, $2=\{1,0\}$, $\cdots$ are the finite ordinals. The set ω of all natural numbers is the first limit ordinal.

An important set theoretic construction generates the real and complex numbers from the natural numbers. The integers are constructed as equivalence classes of natural numbers and the rationals are equivalence classes of integers. The reals are constructed as Dedekind cuts of rationals (or equivalence classes of Cauchy sequences of rationals) and the complex numbers as ordered pairs of reals.

The proper class, On, of all ordinals can be used to order the stages in the construction of a universe of sets. For example, V defined by

$$V_0=0,\quad V_\alpha=P(V_\alpha),$$

$$V=\bigcup_{\alpha<\beta}V_\alpha \text{ if } \alpha \text{ is a limit ordinal},$$

$$V=\bigcup_{\alpha\in On}V_\alpha,$$

where $P(V_\alpha)$ is the set of all subsets (power set) of V_α, is a universe of sets.[10]

A model $\mathbf{M}$ of ZFC (ZF set theory with the axiom of choice) is an ordered pair $\langle M,\mathrm{R}\rangle$, where M is a class and R is a binary relation such that all the axioms of ZFC are true in $\mathbf{M}$ with R interpreted as the membership relation and M the universe of sets in $\mathbf{M}$. A model $\mathbf{M}$ is transitive if M is transitive. $\mathbf{M}$ is standard if R is the the usual membership relation, i.e., if $x\in y\leftrightarrow x\,\mathrm{R}\,y$ for all x,y in M. From now on all ZFC models considered here will be standard transitive models. $\mathbf{V}=\langle V,\in\rangle$ with V defined above is a standard transitive ZFC model.

There exist many nonisomorphic ZFC models, both by the Lowenheim Skolem theorems and by direct construction. However, by Gödel's incompleteness theorem this cannot be proved in ZFC. A stronger set theory, such as Kelly Morse (Quine Morse) set theory[8,11,12] is required. The general type of ZFC model theoretic expressions which are often theorems of ZFC are expressions of the form "If ZFC has a (standard) model then $\cdots$" or "If ZFC is consistent then . . . ".

For any class M and any formula ϕ, $\mathbf{M}\models\phi$ means ϕ is true in $\mathbf{M}$. $\mathbf{M}\models\phi$ stands for a formula in $\mathcal{L}_{\mathrm{ZF}}$, the language of ZF set theory, which is defined inductively as follows[13]:

$$\mathbf{M}\models x\in y\ \leftrightarrow\ x\in\mathbf{M}\text{ and }y\in\mathbf{M}\text{ and }x\,\mathrm{R}\,y,$$

$$\mathbf{M}\models\neg\phi\ \leftrightarrow\ \mathbf{M}\models\phi\text{ is false},$$

$$\mathbf{M}\models\phi\vee\phi'\leftrightarrow\mathbf{M}\models\phi\text{ or }\mathbf{M}\models\phi',$$

$$\mathbf{M}\models\phi\wedge\phi'\leftrightarrow\mathbf{M}\models\phi\text{ and }\mathbf{M}\models\phi',$$

$$\mathbf{M}\models\exists x\phi\ \leftrightarrow\ \exists\,x\in M(\mathbf{M}\models\phi),$$

$$\mathbf{M}\models\forall\,x\phi\ \leftrightarrow\ \forall\,x\in M(\mathbf{M}\models\phi).$$

It is provable in ZFC that there is a formula in ZFC which expresses $\mathbf{M}\models(-)$ for all formulas if and only if M is a set.[14] If M is a proper class, no single formula of ZFC expresses $\mathbf{M}\models(-)$ for all formulas in $\mathcal{L}_{\mathrm{ZF}}$. How-However, under most any restriction on the set of formulas, $\mathbf{M}\models(-)$ is expressible in $\mathcal{L}_{\mathrm{ZF}}$. For example, $\mathbf{M}\models(-)$ restricted to any finite set of formulas or to the set of all formulas with $\leq n$ quantifiers or to formulas with quantifiers restricted to some set are all expressible by a formula in the language of ZF.

A very important concept is that of absoluteness.[15] Absoluteness has to do with properties of mathematical objects as seen from inside and outside various ZFC models. Let $\phi(x)$ be a formula of $\mathcal{L}_{\mathrm{ZF}}$ with x free in ϕ and $\mathbf{M}$ a standard transitive ZFC model. ϕ is $\mathbf{M}$ absolute if

$$\phi(a)\leftrightarrow\mathbf{M}\models\phi(a) \tag{1}$$

holds for all $a\in M$. If ϕ is a sentence of $\mathcal{L}_{\mathrm{ZF}}$, then ϕ is $\mathbf{M}$ absolute if $\phi\leftrightarrow\mathbf{M}\models\phi$ holds.

The meaning of Eq. (1) is the following. The formula ϕ determines a property of sets. A property of sets is $\mathbf{M}$ absolute if for each a in $\mathbf{M}$ a has the property when viewed inside $\mathbf{M}$ if and only if a has the property when viewed outside $\mathbf{M}$. A set is $\mathbf{M}$ absolute if and only if its defining relation is $\mathbf{M}$ absolute.

A formula ϕ is *absolute* if it is absolute for all standard transitive ZFC models. A set is *absolute* if its defining relation is absolute. Note that absoluteness resembles in some ways the physical concept of invari-

ance. In particular, it follows from Eq. (1) that if $\phi(x)$ is absolute, then, for every pair of standard transitive ZFC models **M** and **N**, if $M \subset N$,

$$\mathbf{N} \models \phi(a) \leftrightarrow \mathbf{M} \models \phi(a)$$

holds for all $a \in M$.

Many of the common sets and concepts are absolute. For example, $x=y$, $x \in y$, $x \subset y$, x is a function, x is a ordinal, etc.,[15] are absolute. So is "x is a natural number," "x is a rational number," "x is a real number" (as a Dedekind cut on the rationals), and "x is a complex number." Let ϕ and θ be formulas. If ϕ and θ are absolute, so are $\theta \vee \phi$, $\phi \vee \theta$, $\neg\, \phi$, $\exists\, x \in y(\phi)$, and $\forall\, x \in \in y(\phi)$.[15]

As an example of a property which is not absolute, let $\phi(x)$ be the formula of $\mathcal{L}_{ZF}$ which expresses "x is the set of real numbers" and let **M** be any countable standard transitive ZFC model. Then there is a set R^M such that $\mathbf{M} \models \phi(R^M)$. But outside of **M**, R^M is countable and thus is clearly not the set of R of real numbers outside **M**. Since "y is a real number" is absolute $R^M = R \cap M$. Note also that inside **M** R^M is uncountable (this is another property which is not absolute). There is no contradiction here since all 1–1 maps from ω (the set of all natural numbers or finite ordinals) onto R^M, by means of which one defines the countability of R^M, are outside of **M**.

C. Uniqueness of the scalar field

One first notes that essentially all mathematical physics and most of mathematics refers to *many* Hilbert spaces, groups, algebras, etc., but to only *one* field of complex numbers, only *one* field of real numbers and rational numbers, and only one set of natural numbers. This is evident from references in the literature to *the* real and complex numbers. [It is assumed here that the scalar field for quantum mechanics is a complex number field and not a real or quaternion field.]

The usual axiomatization of the scalar field, as an algebraically closed field of characteristic 0, ACF(0),[3] fails to give this uniqueness. By the Lowenheim Skolem theorems there are many nonisomorphic models of ACF(0).[16] However, this axiomatization gives almost the best possible result for a first order theory since for each uncountable cardinal, λ, all models of ACF(0) of cardinality λ are isomorphic.[16,17]

There are several ways to recover the uniqueness of the complex number field. One way is to require that the field as a model of ACF(0) have cardinality $2^{\aleph_0}$. Then there is only one such field. [From now on, "one," "same," or "unique" will always mean up to isomorphism.]

Another way is to require that the complex number field, as a model of ACF(0), be connected and locally compact. Then there is only one such field.[18,19] Another method is by the direct construction (given earlier) which begins with the empty set. Each of these characterizations gives the same complex number field. Similar considerations apply to the real numbers.

The above discussion can be formalized in ZFC. Then all the relevant theorems become theorems of ZFC and are true in every ZFC model universe. Thus inside each ZFC model universe there exists a unique field of complex numbers of cardinality $2^{\aleph_0}$ and which is connected and locally compact and is built up from the empty set. In particular the complex number field used in physics so far is this unique field in the ZFC model V.

However, the above definitions of the complex number field are not absolute. Thus let **M** and **N** be standard transitive ZFC models and c and d sets such that, inside **M** and inside **N**, c and d are the respective unique complex number fields. In general, outside **M** and **N**, c and d are not isomorphic. If $M \subset N$, then this discussion can be carried on inside N.

III. ZFC MODELS AS MATHEMATICAL UNIVERSES FOR QUANTUM MECHANICS

A. Necessary conditions for a model

From now on the discussion will be restricted to quantum mechanics. Except for Secs. B and C, the further restriction to observables which are questions will be made. Such a restriction is innessential, and it keeps the mathematics as simple as possible while retaining all the essential features.

Let **M** be a standard transitive ZFC model. The following conditions are clearly necessary, but not sufficient if M is to be a suitable mathematical universe for quantum mechanics.

(a): Let $\mathcal{S}$ and $\mathcal{Q}$ be, respectively, collections of state preparation procedures and question measuring procedures. Let $\mathcal{H}_M$ and $B(\mathcal{H}_M)$ be sets such that, inside **M**, $\mathcal{H}_M$ is a Hilbert space, and $B(\mathcal{H}_M)$ is the set of all bounded linear operators over $\mathcal{H}_M$. Then there must exist maps Ψ_M from $\mathcal{S}$ to $B(\mathcal{H})$ and Φ_M from $\mathcal{Q}$ to $B(\mathcal{H})$ such that for all $s \in \mathrm{Dom}\Psi_M$ and $b \in \mathrm{Dom}\Phi_M$, inside **M**, $\Psi_M(s)$ and $\Phi_M(b)$ are, respectively, a density operator and and a projection operator in $B(\mathcal{H}_M)$.

(b): Let $t: \omega \to R_M$ be an increasing function inside **M**. Let (t,s,b) denote the process "carry out state preparation procedure s, then carry out question procedure b on the system so prepared, observe outcome and discard system. Repeat this $s-b$ measurement at calendar times $t(0)$, $t(1)$, ..."

Then for each s, b, and t for which (t,s,b) is an infinite repetition of doing s and b at times $t(0)$, $t(1)\circ\circ\circ$, there exists a sequence ψ_{tsb} which must satisfy the following requirements: (1) $\psi_{tsb} \in M$ and ψ_{tsb} is random, (2) *inside* **M**

$$\bar{M}\psi_{tsb} = \mathrm{Tr}_M(\Psi_M(s)\Phi_M(b)) \tag{2}$$

holds where $\bar{M}\psi$ denotes the limit mean of ψ, (3) for each m the finite outcome sequence obtained by carrying out the first m repetitions of (t,s,b) is given by the first m elements of ψ_{tsb}.

These conditions (a) and (b) are necessary but clearly not sufficient for **M** to be a carrier for the mathematics of quantum mechanics. Condition (a) states the requirement that preparation and question procedures must correspond to states and projection operators in **M**. Note that "$\mathcal{H}$ is a Hilbert space" is not in general **M**

absolute. Thus the object H which is a Hilbert space in **M** is not in general a Hilbert space outside **M**. This will be discussed more later on.

In regard to condition (b), it is clear that there are many processes which correspond to an infinite repetition of doing s and b. Among other things, these processes are distinguished by when each repetition is to be carried out. It is assumed here that the instructions for s do not specify when s is to be carried out and the instructions for b include a time delay relative to preparation procedures. It is assumed that s and b include instructions on where they are to be done. Also inclusion of external fields is suppressed.[20] It should be clear that these assumptions are in some ways arbitrary and can easily be changed or relaxed without affecting the results of this work. The essential part t plays as far as this paper is concerned is as a name or label by means of which different infinite repetitions of s and b are distinguished.

Note that condition (b) requires the mathematical existence of ψ_{tsb} in M. It does not say anything about or require that one actually be able to complete an infinite repetition of doing s and b. All one ever has in hand at any finite time is a finite initial segment of ψ_{tsb} [part (3)] of condition (b).

The question arises why in condition (b) it is required that $\psi_{tsb} \in M$ and Eq. (2) hold in **M**. The reason is that if **M** is to serve as the mathematical universe for quantum mechanics, then **M** is also the universe within which one computes the mean of the outcomes of the first n repetitions of (t,s,b) [i.e., $\bar{M}_n\psi_{tsb}=\sum_{j=0}^{n-1} n^{-1}\psi_{tsb}(j)$] and requires that $\lim_n \bar{M}_n \psi_{tsb} = \mathrm{Tr}_M(\Psi_M(s)\Phi_M(b))$ or Eq. (2) hold. Note, requiring $\psi_{tsb} \in M$ is the same as requiring the existence, in **M** of an infinite set of finite 0–1 sequences $\psi_n \in \{0,1\}^n$ for $n=1,2,\cdots$ such that $m>n$ implies ψ_n is an initial segment of ψ_m. Note that S, Q, Ψ_M, and Φ_M are all outside M.

In condition (b) as given, the requirement that ψ_{tsb} is random is outside **M**, whereas one would expect that if **M** is to be the mathematical universe for quantum mechanics the ψ_{tsb} should be random inside **M**. Then all tests for randomness would be carried out inside **M**.

The point to be made here is that for all the "usual" tests for randomness it makes no difference whether they are carried out inside or outside **M**. The reason is that their description is **M** absolute. This can be stated more precisely as follows: If T is an **M** absolute test for randomness, then, outside **M**, ψ passes T and $\psi \in M$ if and only if inside **M** ψ passes T. Thus with respect, to these tests it is immaterial whether one requires in condition (b) "ψ_{tsb} is random "or" $\mathbf{M} \models \psi_{tsb}$ is random."

However, for strong definitions of randomness, there are "esoteric" tests for randomness which cannot even be defined in **M** if M is sufficiently "small." This point, which will be discussed more in Sec. VI, is why the requirement of randomness in condition (b) is given outside **M**.

B. Hilbert spaces inside the and outside M

As an aid in understanding it is worthwhile to compare some aspects of quantum mechanics based in **M** and quantum mechanics based in the usual intuitive mathematics.

Let s and b be state preparation and question measuring procedures. Then by condition (a), inside **M**, $\Psi_M(s)$ is a density operator in $B(H_M)$ and $\Phi_M(b)$ is a projection operator in $B(H_M)$. For quantum mechanics based on the usual mathematics one also has by condition (a) maps Ψ and Φ such that $\Psi(s)$ and $\Phi(b)$ are respective density and projection operators in some $B(H)$.

Now one would like to know the relationship between Ψ and Ψ_M and Φ and Φ_M. In particular one would like to have the result that

$$\mathrm{Tr}_M(\Psi_M(s)\Phi_M(b)) = \mathrm{Tr}(\Psi(s)\Phi(b)) \tag{3}$$

holds outside M for all s and b in the respective domains of Ψ_M and Φ_M.

The main goal of this and the next sections is to show that this is always possible. That is, we shall show that if $\mathbf{M} \models B(H_M)$ is the set of all bounded linear operators over the Hilbert space H_M, then there is a Hilbert space H and operator algebra $B(H)$ outside M as well as a natural correspondence from H_M into H and from $B_M(H_M)$ into $B(H)$. We shall also show that if "T is an operator in $B_M(H_M)$" is true in **M**, then the operator in $B(H)$ which corresponds to T has the same eigenvalues as T. The main goal of this section is to prove the following theorem.

Theorem 1: Let **M** be a standard transitive ZFC model and let H_M be a set such that $\mathbf{M} \models H_M$ is a Hilbert space. Then outside **M** there exists a set H and a map U_M such that H is a Hilbert space and U_M is an isometric monomorphism of H_M into H.

The proof will be given as a series of Lemmas. First the properties of H_M outside **M** must be established. Let D be a subset of the complex number field C such that D is closed under multiplication, addition, and complex conjugation. A set J is a D pre-Hilbert space if J satisfies the axioms of a pre-Hilbert space with the scalars restricted to D.

Lemma 2: Let **M** and H_M be as in Theorem 1 and C_M the set of all complex numbers inside **M**. Then, outside **M**, H_M is a C_M pre-Hilbert space.

Proof: H_M is a set, $x \in H_M$, $x=y$, and $+$ is a binary function $H_M \times H_M \to H_M$ are all **M** absolute.[15] Thus $x+y=z$ holds if and only if $\mathbf{M} \models x+y=z$ for all $x,y,z \in H_M$. Thus one has $x+y=y+x \leftrightarrow \mathbf{M} \models x+y=y+x$ and $x+(y+z)=(x+y)+z \leftrightarrow \mathbf{M} \models x+(y+z)=(z+y)+z$ and $x+0=x \leftrightarrow \mathbf{M} \models x+0=x$. Since $\mathbf{M} \models H_M$ is a Hilbert space, $+$ is associative and commutative on H_M, and 0 is the additive identity element outside **M**.

Similar arguments apply to scalar multiplication of vectors. Note that multiplication and addition of complex numbers themselves is absolute so $\mu \circ \gamma \in C_M \leftrightarrow \mathbf{M} \models \mu \circ \gamma \in C_M$, $\gamma^* = \mu \leftrightarrow \mathbf{M} \models \gamma^* = \mu$, etc. Thus, outside **M**, C_M is closed under multiplication, addition, and complex conjugation.

Finally "the scalar product $(\ ,\ )$ is a binary function from $H_M \times H_M$ to C_M" is **M** absolute. Use of this with the

above gives the result that $(x,y)=(y,x)^*$, $(x,y+z)=(x,y)+(x,z)$, $(x,\lambda y)=\lambda(x,y)$, and $\|x\|^2=(x,x)>0$ unless $x=0$ are all **M** absolute. Since H_M is a Hilbert space inside **M**, (,) satisfies the requisite axioms outoutside **M**. QED

Note that since both C_M and H_M are complete inside **M**, they are **M** complete outside **M**. However, they are not complete outside **M**. By C_M (or H_M) being **M** complete is meant that each Cauchy sequence ψ of elements of C_M (or H_M) such that ψ lies in M, converges to an element of C_M (or H_M).

To construct a Hilbert space H from H_M, it is necessary to complete H_M. The literature results cannot be taken over directly because H_M is defined only over C_M and not over C. However, the changes needed are minor since C_M is dense in C [C_M contains all the rational complex numbers].

To construct H, one can proceed as follows[21]: Let C_M^ω and H_M^ω denote the sets of all Cauchy sequences, outside **M**, of scalars in C_M and vectors in H_M. Thus $f\in C_M^\omega$ (or H_M^ω) implies that for each n, $f(n)\in C_M$ [or $f(n)\in H_M$] and $\forall n\exists m\forall k>m\forall l>m|f(k)-f(l)|<2^{-n}$ [or $\forall n\exists m\forall k>m\ \forall l>m\|f(k)-f(l)\|<2^{-n}$]. Note that there are many such f which are not in M.

Lemma 3: (1) f is a Cauchy sequence of elements of C_M (or H_M) is **M** absolute; (2) $f\in C_M^\omega$ or $f\in H_M^\omega$ implies f is bounded.

Proof: (1) By Lemma 2 and the absoluteness of arithmetic operations on the natural numbers $\|f(k)-f(l)\|<2^{-n}$ and $|f(k)-f(l)|<2^{-n}$ are **M** absolute. Since $\omega=$ set of all natural numbers is **M** absolute, the definitions of Cauchy convergence for any $f\in H_M^\omega\cap M$ or $f\in C_M^\omega\cap M$ are **M** absolute [f is an infinite sequence of elements of H_M or of C_M is **M** absolute]. (2) Since C is complete and $C_M=C\cap M$, $f\in C_M^\omega$ implies that $\lim_n f(n)$ exists in C and that $\sup_n|f(n)|$ exists in C. For $f\in H_M^\omega$ one has $\|f(k)\|-\|f(l)\|<\|f(k)-f(l)\|$. Thus the sequence $\{\|f(k)\|\ |\ k\in\omega\}$ is Cauchy convergent and, by the above, $\sup_n\|f(n)\|$ exists in C. QED

Let [0] be the set of all sequences of H_M^ω which converge to 0. That is $f\in[\mathbf{0}]$ if $\forall n\exists m\forall k>m\|f(k)\|<2^{-n}$. Define $f\sim g$ if $f-g\in[\mathbf{0}]$. Clearly $\sim$ is an equivalence relation. Let $[f]=\{g\,|\,g\sim f\}$ and define H to be $H_M^\omega/\sim$, the set of all equivalence classes of elements of H_M^ω.

The various operations on H_M are lifted to H as follows: $f+g$ is defined by $(f+g)(n)=f(n)+g(n)$ for all $n\in\omega$ and $f,\ g\in H_M^\omega$. For each $\alpha\in C_M^\omega$ and $f\in H_M^\omega$ one similarly defines $\alpha\circ f$ by $(\alpha\circ f)(n)=\alpha(n)f(n)$ for each n. By Lemma 2, both definitions are possible.

One then defines addition and scalar multiplication on H by

$$[f]+[g]=[f+g], \tag{4}$$

$$\gamma\circ[f]=[\alpha_\gamma\circ f], \tag{5}$$

for all f and g in H_M^ω and α_γ in C_M^ω such that $\lim_n\alpha_\gamma(n)=\gamma$. Since C_M is dense in C, for every γ in C there exists such an α_γ in C_M^ω. By standard arguments one shows that Eqs. (4) and (5) are well defined on H by showing that if $f\sim f'$, $g\sim g'$, and $\alpha\sim\alpha'$, then $f+g\sim f'+g'$ and $\alpha\cdot f\sim\alpha'\cdot f'$ for all f,f', g,g' in H_M^ω and α and α' in C_M^ω.

By use of Lemma 2 one defines a scalar product on H_M^ω as

$$(f,g)=\lim_n(f(n),g(n)).$$

It is easily shown by polar decomposition and Lemma 3 that the limit exists, and, if $f\sim f'$ and $g\sim g'$, then $(f,g)=(f',g')$. Thus one defines a scalar product on H by

$$([f],[g])=(f,g) \tag{6}$$

for all f and g in H_M^ω.

Lemma 4: H is a Hilbert space over C.

Proof: One first shows that H is a pre-Hilbert space over C. Let $\gamma,\ \mu\in C$, then $(\gamma+\mu)\cdot[f]=[(\alpha_\gamma+\alpha_\mu)f]$, where $\alpha_\gamma\in[\gamma]$, $\alpha_\mu\in[\mu]$. We define $[\gamma]=\{\alpha\,|\,\alpha\in C_M^\omega$ and $\lim_n\alpha(n)=\gamma\}$. Now $((\alpha_\gamma+\alpha_\mu)\cdot f)(n)=(\alpha_\gamma+\alpha_\mu)(n)\circ f(n)$ $=$ (Lemma 2) $\alpha_\gamma(n)\cdot f(n)+\alpha_\mu(n)\circ f(n)$. So $[(\alpha_\gamma+\alpha_\mu)\circ f]=[\alpha_\gamma\cdot f+\alpha_\mu\cdot f]=[\alpha_\gamma\circ f]+[\alpha_\mu\cdot f]=\gamma\cdot[f]+\mu\cdot[f]$. The proof of the other pre-Hilbert space axioms is handled in a similar fashion and is left to the reader. It remains to show that H is complete. Since the proof given by Yosida[21] can be used without modification, it is also left to the reader. QED

Theorem 1 can now be proved. Let **M** be a standard transitive ZFC model. Let H_M be such that $\mathbf{M}\models H_M$ is a Hilbert space. By Lemma 3, outside **M**, H_M is a C_M pre-Hilbert space. By Lemmas 3, and 4 , $H=H_M^\omega/\sim$ is a Hilbert space constructed from H_M.

For each element $a\in H_M$ let $\bar a$ denote the constant a sequence and $[\bar a]$ the corresponding element of H. Let U_M be the map: $H_M\to H$ defined by $U_Ma=[\bar a]$. By construction of the scalar product, Eq. (6), on H, $\|a\|=\|[\bar a]\|$ so that U_M is isometric. Finally $a-b\neq0\to[\bar a]-[\bar b]\neq[\mathbf{0}]$ so that U_M is 1–1. By Eq. (4) $U_M(a+b)=U_M(a)+U_M(b)$. By Eq. (5), $U_M(\gamma a)=[\overline{\gamma a}]=[\bar\gamma\circ\bar a]=\gamma\circ[\bar a]=\cdot\ \gamma\circ U_M(a)$ for all $\gamma\in C_M$ (is the constant γ sequence.) Thus U_M is an isometric monomorphism and the theorem is proved. QED

This theorem thus insures the existence of a Hilbert space $H_M^\omega/\sim$ given an H_M which is a Hilbert space in **M**, when no further conditions on H_M are present in **M**. If further conditions are present, then different constructions must be used. For example, if $\mathbf{M}\models H_M=L^2(R,\beta(R),\mu)_M$, then the corresponding space outside **M** is $L^2(R,\beta(R),\mu)$ which is not $H_M^\omega/\sim$.

Finally one notes the following properties of H and H_M: Let B be a subset of H_M inside **M** and let $\mathbf{B}=\{[\bar a]\,|\,a\in B\}$. Then

Corollary 5: (1) $\mathbf{M}\models B$ is an orthonormal basis set for $H_M\to\mathbf{B}$ is an orthonormal basis set for H. (2) $\mathbf{M}\models H_M$ is separable. H is separable.

Proof: (1) By Theorem 1 and the construction of **B**, $\mathbf{M}\models$(the elements of B are orthonormal and linearly independent) $\leftrightarrow$ the elements of **B** are orthonormal and linearly independent. Let F in **M** be, inside **M**, the set of all finite linear combinations of elements of B and let $\mathbf{F}=\{[\bar a]\,|\,a\in F\}$. One must show that **F** is dense in H. First for any $a\in H_M$ and any n let $c\in F$ be such that $\mathbf{M}\models\|a-c\|<2^{-n}$. Then outside **M**, $\|a-c\|<2^{-n}$ and $\|[\bar a]-[\bar c]\|<2^{-n}$ where $[c]\in\mathbf{F}$.

For the general element $[f]$ of H, one has the fol-

lowing: For each n let m_n be such that for all $k \geq m_n$, $\| f(k) - f(m_n) \| < 2^{-n}$. Since f is Cauchy, such an m_n exists. Then $\| [f] - [\overline{f(m_n)}] \|^2 = \lim_p (f(p) - f(m_n),\ f(p) - f(m_n)) < (2^{-n})^2$ or $\| [f] - [\overline{f(m_n)}] \| < 2^{-n}$. By hypothesis and the above proof, there is an a in F such that $\| a - f(m_n) \| < 2^{-n}$. Thus $\| [f] - [a] \| < \| [f] - [\overline{f(m_n)}] \| + \| [\overline{f(m_n)}] - [a] \| < 2^{-n} + 2^{-n}$. Since n is arbitrary, $\mathbf{F}$ is dense in H.

(2) By hypothesis $\mathbf{M} \models \exists$ orthonormal basis set B for H_M which is countably infinite. The latter means that there exists in $\mathbf{M}$ a bijection h from ω onto B. Clearly then, outside $\mathbf{M}$, $\mathbf{B}$ is countably infinite, or from (1) H is separable. QED

It should be noted that in general the implication in part (2) cannot be reversed. For example, if H_M is non-separable inside $\mathbf{M}$, and M is countable, then, by part (1) above, H is separable.

C. $B(\mathcal{H})$ inside and outside of M

For any Hilbert space H let $B(H)$ denote the set of bounded linear operators over H. A first goal is to prove a theorem for operators which corresponds to Theorem 1 for H and H_M.

One first defines a operator T on H_M outside $\mathbf{M}$ to be C_M bounded linear if $T(a+c) = Ta + Tc$, $T(\gamma a) = \gamma \cdot (Ta)$ for all a and c in H_M and for all $\lambda \in C_M$, and there exists a $\mu \in C_M$ such that $|(a, Tc)| < \mu |(a, c)|$ for all $a,\ c \in H_M$. Then one has

Lemma 6: Let $\mathbf{M}$ be a standard transitive ZFC model. Then

(1) $\mathbf{M} \models T$ is a bounded linear operator on $H_M \rightarrow$ outside $\mathbf{M}$, T is a C_M bounded linear operator on H_M.

(2) The algebraic operations on $B(H_M)$ are $\mathbf{M}$ absolute.

Proof: (1) T is a map $H_M \rightarrow H_M$ is $\mathbf{M}$ absolute. By Lemma 2, addition on H_M and multiplication by a scalar in C_M are $\mathbf{M}$ absolute. Thus $T(a+c) = b \rightarrow \mathbf{M} \models T(a+c) = b \rightarrow \mathbf{M} \models Ta + Tc = b \rightarrow Ta + Tc = b$. For each scalar $\lambda \in C_M$, $T(\lambda a) = b \rightarrow \mathbf{M} \models T(\lambda a) = b \rightarrow \mathbf{M} \models \lambda(Ta) = b \rightarrow \lambda \circ (Ta) = b$. So T is C_M linear.

Also by the proof of Lemma 2 $(a, Tb) = r \rightarrow M \models (a, Tb) = r$. Thus, if x is a bound for T inside $\mathbf{M}$, x is a bound for T outside $\mathbf{M}$ and T is bounded.

(2) We must show that the defining operations for multiplication, addition, scalar multiplication, and the adjoint are $\mathbf{M}$ absolute.[15] By Lemma 2 and part (1) above:

(a) $\mathbf{M} \models W = S + T \leftrightarrow \mathbf{M} \models \forall a \in H_M (Wa = Sa + Ta) \leftrightarrow \forall a \in H_M\ (Wa = Sa + Ta) \leftrightarrow W = S + T$

(b) Let $\lambda \in C_M$. Then $\mathbf{M} \models W = \lambda T \leftrightarrow \mathbf{M} \models \forall a \in H_M\ (Wa = \lambda(Ta)) \leftrightarrow \forall a \in H_M\ (Wa = \lambda(Ta)) \leftrightarrow W = \lambda \circ T$.

(c) $\mathbf{M} \models W = S \circ T \leftrightarrow \mathbf{M} \models \forall a \in H_M (W(a) = S(T(a))) \leftrightarrow \forall a \in H_M\ (W(a) = S(T(a))) \leftrightarrow W = S \circ T$. Here we have used the fact that composition of maps is absolute.

(d) $\mathbf{M} \models S = T^\dagger \leftrightarrow \mathbf{M} \models \forall\ a, b \in H_M (Ta, b) = (a, Sb) \leftrightarrow \forall\ a, b \in H_M\ (Ta, b) = (a, Sb) \leftrightarrow S = T^\dagger$. QED

For each T in $B(H_M)$ define $T' : H_M^\omega \rightarrow H_M^\omega$ and $\mathbf{T} : H \rightarrow H$ by $(T'f)(n) = T(f(n))$ for each n and

$$\mathbf{T}[f] = [T'f] \tag{7}$$

for each f in H_M^ω. By standard arguments one shows that if $f \sim f'$, then $T'f \sim T'f'$.

Lemma 7: For each $T \in B(H_M)$

(1) $\mathbf{T}$ is a bounded linear operator on H,

(2) $\mathbf{M} \models (\lambda = \| T \|) \rightarrow \lambda = \| \mathbf{T} \|$.

Proof: (1) By Eqs. (7) and (4), $\mathbf{T}([f] + [g]) = \mathbf{T}([f+g]) = [T'(f+g)] = (\text{Lemma 6})\ [T'f + T'g] = [T'f] + [T'g] = \mathbf{T}[f] + \mathbf{T}[g]$. Let $\gamma \in C$ and $\boldsymbol{\alpha}_\gamma \in C_M^\omega$ with $\lim \alpha_\gamma = \gamma$. Then by Eqs. (7) and (5), $\mathbf{T}(\gamma \cdot [f]) = \mathbf{T}([\boldsymbol{\alpha}_\gamma f]) = [T' \boldsymbol{\alpha}_\gamma f] = (\text{Lemma 6})\ [\boldsymbol{\alpha}_\gamma \cdot T'f] = \gamma \cdot [T'f] = \gamma \cdot \mathbf{T}([f])$. So $\mathbf{T}$ is linear on H. Let $\gamma \in C_M$ be a bound for T inside $\mathbf{M}$. Then, by the the proof of Lemma 6, γ is a bound for T outside and $\| \mathbf{T}[f] \| = \| [T'f] \| = [\lim_n \| (T'f)(n) \|^2]^{1/2} \leq \gamma [\lim_n \| f(n) \|^2]^{1/2} = \gamma \| [f] \|$. So γ is a bound for T.

(2) Suppose there exists a $\gamma \in C$ with $\gamma < \lambda$ such that $\gamma =$ least upper bound for T. Then since C_M is dense in C there exists a $\delta \in C_M$ with $\gamma < \delta < \lambda$. So δ is an upper bound for T. Then for all $a \in H_M$ one has, outside $\mathbf{M}$, $\| Ta \| = (T\bar{a}, T\bar{a})^{1/2} = \| \mathbf{T}[\bar{a}] \| \leq \delta \| [\bar{a}] \| = \delta \| a \|$ so δ is an upper bound for T outside M. Since $\| Ta \| < \delta \| a \|$ is $\mathbf{M}$ absolute, $\mathbf{M} \models (\| Ta \| < \delta \| a \|$ for all $a \in H_M$ and $\delta < \lambda)$ which contradicts λ being the least upper bound of T inside M. Thus λ is also the least upper bound of T. QED

These results can be combined into the following theorem which is the main goal of this section.

Theorem 8: Let $\mathbf{M}$ be a standard transitive ZFC model and let $B(H_M)$ be the set of all bounded linear operators over a Hilbert space H_M inside $\mathbf{M}$. Let H be as constructed from H_M in Theorem 1. Then the map $V_M : B(H_M) \rightarrow B(H)$ given by $V_M T = \mathbf{T}$ [Eq. (7)] is an isometric monomorphism of $B(H_M)$ into $B(H)$.

Proof: By Lemma 7, the map V_M is into $B(H)$ and is isometric. Let S and $T \in B(H_M)$ with $\mathbf{M} \models S \neq T$. By absoluteness $S \neq T$ so that there is an $a \in H_M$ such that $Sa \neq Ta$ or $[\overline{Sa}] \neq [\overline{Ta}]$ which gives $\mathbf{S}[\bar{a}] \neq \mathbf{T}[\bar{a}]$ or $\mathbf{S} \neq \mathbf{T}$. So the map V_M is 1–1. Let $\mathbf{M} \models W = T \circ S$ with $W, T, S, \in B(H_M)$. Then for each $[f] \in H$. $\mathbf{W}[f] = [W'f] = [(T \cdot S)'f] = [T'(S'f)] = \mathbf{T} \cdot [S'f] = \mathbf{T} \cdot \mathbf{S} \cdot [f]$ or $V_M(T \circ S) = V_M(T) \cdot V_M(S)$, where Lemma 6 has been used. Let $\mathbf{M} \models W = \lambda \cdot T$ with $\lambda \in C_M$. Then $\mathbf{W}[f] = [W'f] = [\bar{\lambda} \cdot T'f] = \lambda \cdot [T'f] = \lambda \cdot \mathbf{T}[f]$ or $V_M(\lambda T) = \lambda \cdot V_M(T)$(Lemma 6). Let $W, S, T \in B(H_M)$ be such that $\mathbf{M} \models W = S + T$. Then (Lemma 6) $\mathbf{W}[f] = [W'f] = [(S' + T')f] = [S'f + T'f] = \mathbf{S}[f] + \mathbf{T}[f] = (\mathbf{S} + \mathbf{T})[f]$ or $V_M(S+T) = V_M(S) + V_M(T)$. Finally let $T^\dagger$ be the adjoint of T inside $\mathbf{M}$. Then by Lemma 6, $(Ta, b) = (a, T^\dagger b)$ for all $a,\ b \in H_M$ and $([f],\ (\mathbf{T})^\dagger\ [g]) = (\mathbf{T}[f], [g]) = ([T'f], [g]) = \lim_n (T(f(n))g(n)) = \lim_n (f(n),\ T^\dagger(g(n))) = (f, T^\dagger g) = ([f],\ (\mathbf{T}^\dagger)[g])$. Since this holds for every $[f],\ [g] \in H$, $(\mathbf{T})^\dagger = (\mathbf{T})^\dagger$ or $(V_M T) = V_M(T^\dagger)$. QED

Corollary 9:

(1) $\mathbf{M} \models T$ is a projection operator $\rightarrow V_M T$ is a projection operator.

(2) $\mathbf{M} \models T$ unitary $\rightarrow V_M T$ unitary.

(3) $\mathbf{M} \models T$ self-adjoint $\rightarrow V_M T$ self-adjoint.

(4) $\mathbf{M} \models T$ is a density operator $\rightarrow V_M T$ is a density operator.

Proof: By Theorem 8:

(1) $\mathbf{M} \models T^2 = T \rightarrow (V_M T)^2 = V_M T$.

(2) $\mathbf{M} \models T^\dagger T = T T^\dagger = 1 \rightarrow (V_M T)^\dagger (V_M T) = (V_M T)(V_M T)^\dagger = 1$.

(3) $\mathbf{M} \models T = T^\dagger \rightarrow V_M T = (V_M T)^\dagger$.

(4) Let B be any complete basis set for H_M inside $\mathbf{M}$ and let Tr denote the trace operation on $B(H)$ defined by $\mathrm{Tr}\mathbf{T} = \sum_{x \in D} (Tx, x)$, where D is any complete orthonormal basis set on H. Then by Corollary 5 with $\mathbf{B} = \{[\bar{a}] \mid a \in B\}$ $\mathbf{M} \models \mathrm{Tr}_M T = 1 \rightarrow \mathbf{M} \models \sum_{a \in B} (Ta, a) = 1 \rightarrow \sum_{a \in B} ((\mathbf{T}[\bar{a}], [\bar{a}]) = 1)) \rightarrow \mathrm{Tr}\mathbf{T} = 1$. Finally, $\mathbf{M} \models 0 \leq (Ta, a) \leq 1 \rightarrow 0 \leq (\mathbf{T}[\bar{a}], [\bar{a}]) \leq 1$ for all $a \in H_M$. But $\{[\bar{a}] \mid a \in H_M\}$ is dense in H so $0 \leq (\mathbf{T}[f], [f]) \leq 1$ for all $[f] \in H$. QED

Note that by the above result the trace operation Tr on $B(H)$ is related to the trace operation Tr_M on $B(H_M)$ inside $\mathbf{M}$ by

$$\mathrm{Tr}\mathbf{T} = \mathrm{Tr}_M T. \tag{8}$$

Finally, we give some aspects of the relationship between eigenvalues of T and of $\mathbf{T}$.

Lemma 10:

(1) $\mathbf{M} \models \lambda$ is an eigenvalue of $T \rightarrow \lambda$ is an eigenvalue of $\mathbf{T}$.

(2) $\mathbf{T}[\bar{a}] = \lambda[\bar{a}]$ for some $a \in H_M \rightarrow \mathbf{M} \models \lambda$ is an eigenvalue of T.

Proof: (1) By Lemmas 6 and 2, $\mathbf{M} \models Ta = \lambda a \rightarrow Ta = \lambda a \rightarrow \overline{Ta} = \overline{\lambda a} \rightarrow [\overline{Ta}] = [\overline{\lambda a}] \rightarrow \mathbf{T}[\bar{a}] = \lambda \circ [\bar{a}]$.

(2) $\mathbf{T}[\bar{a}] = \lambda[\bar{a}] \rightarrow [\overline{Ta}] = [\alpha_\lambda \cdot \bar{a}]$ where $\alpha_\lambda \in C_M^\omega$ is such that $\lim \alpha_\lambda = \lambda$. This implies there is a Cauchy sequence $h \in H_M^\omega$ with $h \sim \bar{a}$ and $\overline{Ta} = \alpha_\lambda h$. Taking the strong limit (in H_M outside M), gives $Ta = \text{s-lim} \alpha_\lambda h = \text{s-lim} \alpha_\lambda a = \lambda \circ a$. Thus $\lambda \in C_M$ as $\text{s-lim} \alpha_\lambda a$ exists in H_M. By Lemmas 2 and 6, $Ta = \lambda a \rightarrow \mathbf{M} \models Ta = \lambda a$. QED

It is worth pointing out some ways in which operators in $B(H)$ which are outside the range set of V_M (Theorem 8) differ from those in $V_M B(H_M)$. For simplicity projection operators only are considered. Let $\mathbf{P} = V_M(P)$ and U be a unitary operator in $B(H)$ which is not in $V_M B(H_M)$ and does not commute with $\mathbf{P}$. Then in general $U\mathbf{P}U^\dagger$ is not in the range of V_M.

A more interesting difference is as follows: Let B be a set which, inside $\mathbf{M}$, is an orthonormal basis set which spans H_M with H_M separable. Then, inside $\mathbf{M}$, to each subset S of B, one can associate the projection operator P_S with S spanning $P_S H$. Correspondingly $V_M(P_S)$ (or $\mathbf{P}_S$) is a projection operator in $B(H)$ with $\mathbf{P}_S H$ spanned by $\mathbf{S} = \{[\bar{a}] \mid a \in S\}$ which is a subset of $\mathbf{B}$.

The point to be made is that there are many subsets of B and of $\mathbf{B}$ which do not lie in $\mathbf{M}$. To each such subset S_1 of B there corresponds a projection operator $P_{\mathbf{S}_1}$ in $B(H)$ such that $P_{\mathbf{S}_1} H$ is spanned by $\mathbf{S}_1$ which is a subset of $\mathbf{B}$. Clearly, for each such S_1, $P_{\mathbf{S}_1}$ is not in the range of V_M as it does not correspond to any operator in $B(H_M)$. Furthermore, there are a great many such operators. In particular, if M is countable, there are countably many such operators in the range of V_M as, outside $\mathbf{M}$, B has only countably many subsets which may be in M. However, there are $2^{\aleph_0}$ subsets of B outside $\mathbf{M}$ and thus $2^{\aleph_0}$ operators of the type $P_{\mathbf{S}_1}$ which are outside the range of V_M. Note that for all such operators outside the range of V_M both S_1 and $B - S_1$ are infinite sets. If either S_1 is finite or $B - S_1$ is finite, then S_1 lies in M and $P_{\mathbf{S}_1}$ lies in the range of V_M.

Finally, one notes that the construction of H and $B(H)$ outside $\mathbf{M}$ can be placed inside $\mathbf{N}$, where $\mathbf{N}$ is any standard transitive ZFC model for which $M \subset N$. The reason is that all properties of sets which were required to be $\mathbf{M}$ absolute are, in fact, absolute for all standard transitive ZFC models N for which $M \subset N$.

For example, Lemma 2 becomes, "Let $\mathbf{M}, \mathbf{N}$ be two standard transitive ZFC models with $M \subset N$. Then $\mathbf{M} \models H_M$ is a Hilbert space and C_M the set of complex numbers $\rightarrow \mathbf{N} \models H_M$ is a C_M pre-Hilbert space." Theorem 1 is changed by replacing "outside $\mathbf{M}$" by "inside $\mathbf{N}$" with $\mathbf{M}$ and $\mathbf{N}$ as above. Similar changes occur in Corollary 5, Lemmas 6 and 7, Theorem 8, and Corollary 9 and Lemma 10.

D. Quantum mechanics inside and outside of M

We now return to the question raised at the beginning of subsection B regarding the relationship between the maps Ψ_M and Ψ and Φ_M and Φ. Ψ and Φ are the maps given by condition (a) for quantum mechanics whose mathematics is based in the usual intuitive set theory (denoted by QM). Note that $\mathrm{QM} \equiv \mathrm{QM}_V$, where V is the real universe of ZF sets. [Condition (a) for (QM) has been extensively developed by Ekstein.[20]]

First the natural assumption will be made that every state preparation procedure s and question procedure b in quantum mechanics based on $\mathbf{M}$ (denoted by QM_M) is respectively a state preparation procedure and question measuring procedure for QM. That is, it is assumed that $\mathrm{Dom}\Psi_M \subseteq \mathrm{Dom}\Psi \subseteq \mathcal{S}$ and $\mathrm{Dom}\Phi_M \subseteq \mathrm{Dom}\Phi \subseteq \mathrm{Dom}\Phi \subseteq \mathcal{Q}$.

By the results of the last two sections one can give a natural correspondence by requiring Ψ, Ψ_M, Φ_M, and Φ to satisfy

$$\Psi(s) = V_M(\Psi_M(s)) \tag{9}$$

for all s in the domain of Ψ_M and

$$\Phi(b) = V_M(\Phi_M(b)) \tag{10}$$

for all $b \in \mathrm{Dom}\Phi_M$. Here V_M is the isometric monomorphism of Theorem 9. By Corollary 10, $V_M(\Psi_M(s))$ is a density operator and $V_M(\Phi_M(b))$ is a projection operator in $B(H)$, where H is related to H_M by Theorem 1 and H_M is the Hilbert space inside $\mathbf{M}$ for which the range of Ψ_M and Φ_M lies in $B(H_M)$.

By Eq. (8) one has from Eqs. (9) and (10) that

$$\mathrm{Tr}(\Psi(s)\Phi(b)) = \mathrm{Tr}_M(\Psi_M(s)\Phi_M(b)) \tag{3}$$

holds for all $s \in \mathrm{Dom}\Psi_M$ and $b \in \mathrm{Dom}\Phi_M$ or that Eq. (3) is satisfied. Thus Ψ, Ψ_M, Φ, and Φ_M satisfy the natural requirement that $\Psi(s)$ corresponds naturally to $\Psi_M(s)$ and $\Phi(b)$ to $\Phi_M(b)$ and the expectation values for measuring $\Phi_M(b)$ on a system prepared in state $\Psi_M(s)$ in QM_M is the same as the expectation value for measuring $\Phi(b)$ on a system prepared in state $\Psi(s)$ in QM.

Furthermore, let p_{sb} denote, inside **M**, the probability measure on $\mathcal{B}(\{0,1\})$, the four element set of all subsets of $\{0,1\}$, which is generated from s and b by Ψ_M and Φ_M, by

$$p_{sb}(\{1\}) = \mathrm{Tr}_M(\Psi_M(s)\Phi_M(b)). \tag{11}$$

Then by the above p_{sb} is also assigned to s and b, by Ψ and Φ. Note that "p_{sb} is a probability measure on $\mathcal{B}(\{0,1\})$" is absolute [*proof*: x is a real number (Dedekind cut) between 0 and 1 is absolute as is $x+y=1$ for x,y real.] Then, outside **M**, p_{sb} is also a probability measure on $\mathcal{B}(\{0,1\})$. As a result the statistics for a single measurement of $\Psi_M(s)$ on $\Phi_M(b)$ in QM_M are the same as the statistics for a single measurement of $\Psi(s)$ on $\Phi(b)$ in QM.

One can also apply a theorem of Halmos[22] both inside **M** and outside **M** to generate the unique product measure $P^M_{sb} = \overset{M}{\otimes} p_{sb}$ and $P_{sb} = \otimes\, p_{sb}$ inside and outside **M** respectively. By condition *b*) these measures gives the statistics of any infinite repetition of doing s and b inside and outside of M. The measures P_{sb} and P^M_{sb} are also related by a natural correspondence. This correspondence requires the coding of Borel subsets of $\{0,1\}^\omega$ into the set of all infinite sequences of natural numbers[23,24] and will be gone into more in the next paper.

In general there may be preparation and observation procedures in QM which are not in QM_M. That is the containments $\mathrm{Dom}\,\Psi_M \subseteq \mathrm{Dom}\Psi$ and $\mathrm{Dom}\Phi_M \subseteq \mathrm{Dom}\Phi$ may be proper. This is particularly true if one admits, as procedures, limits of sequences of other procedures. This is entirely consistent with the above because, as discussed in subsection C, the monomorphism V_M of Theorem 9 is into and not onto, and there are many probability measures on $\mathcal{B}(\{0,1\})$ outside **M**, which do not exist inside **M**.

IV. ELIMINATION OF $\mathbf{M}_0$

In this section it is shown that for strong definitions of randomeness, such as the one used elsewhere,[2] the minimal ZFC model $\mathbf{M}_0$ cannot be a carrier for the mathematics of quantum mechanics.

The definition of randomness used here is as follows: A probability measure P on $\mathcal{B}(\{0,1\}^\omega)$ is *correct* for an infinite 0,1 sequence ψ if for all Borel subsets B of $\{0,1\}^\omega$, which are definable from P, if $PB=1$, then $\psi \in B$. A set B is *definable from* P if there is a formula $Q(x,\mathbf{P})$ in the language $\mathcal{L}_{\mathrm{ZF}}$ of set theory to which a name $\mathbf{P}$ for P has been added such that $\forall x(x \in B \leftrightarrow Q(x,\mathbf{P}))$ holds. A sequence ψ is *random* if there exists a product measure P which is correct for ψ. If P is correct for ψ, one also says that ψ is random for P.

This definition accepts as random the constant 1 sequence and the constant 0 sequence. These can easily be excluded by requiring P to be nonatomic [$P\{\phi\}=0$ for each singleton subset $\{\phi\}$ of $\{0,1\}^\omega$] or, equivalently, $0 < p(\{1\}) < 1$. Such a restriction, however, makes the statement of condition (b) more cumbersome.

Note that from the absoluteness of "p_{sb} [Eq. (11)] is a probability measure and $\bar{M}\psi = p_{sb}(\{1\})$" one has that if ψ is random and inside **M**, $\bar{M}\psi = p_{sb}(\{1\})$, then outside **M**, P_{sb} is correct for ψ where $P_{sb} = \otimes\, p_{sb}$ is constructed outside **M**.

For any set B let $\mathbf{B} = \langle B, \in\rangle$. Define $Df(\mathbf{B})$ to be the set of all subsets of B which are $\mathcal{L}_{\mathrm{ZF}}$ definable from elements of B inside B. That is $Df(\mathbf{B}) = \{C \mid$ for some n there exists a formula Q in $\mathcal{L}_{\mathrm{ZF}}$ with $n+1$ free variables such that $\exists a_1 \cdots a_n \in B \forall a\, (a \in C \leftrightarrow \mathbf{B} \models Q(a, a_1 \cdots a_n))\}$. For each ordinal β define the structure $\mathbf{L}_\beta = \langle C_\beta, \in\rangle$ inductively by

$$L_0 = 0, \quad L_{\beta+1} = Df(\mathbf{L}_\beta), \quad L_\beta = \bigcup_{\gamma<\beta} L_\gamma, \tag{12}$$

if β is a limit ordinal. Then[1,14] $L = \cup_{\beta \in On} L_\beta$, where On is the class of all ordinals, is Gödel's constructible universe and **L** is the smallest standard transitive ZFC model that contains all the ordinals.

Let δ be the smallest ordinal for which there exists a standard transitive ZFC model **M** such that $\delta \notin M$. Define M_0 by

$$M_0 = \bigcup_{\beta<\delta} L_\beta. \tag{13}$$

Then $\mathbf{M}_0 = \langle M_0, \in\rangle$ is the unique minimal standard transitive ZFC model.[1,25] That is, for any standard transitive ZFC model **N**, $M_0 \subseteq N$.

$\mathbf{M}_0$ has the following properties: M_0 is a countable set and M_0 is definable[1,26]. That is, there is a formula q in $\mathcal{L}_{\mathrm{ZF}}$ such that $\forall a(a \in M_0 \leftrightarrow q(a))$. Also each element of M_0 is definable inside **M**. That is, for each $a \in M_0$ there is a formula q in $\mathcal{L}_{\mathrm{ZF}}$ such that[1,25,26] $\mathbf{M}_0 \models \forall y(y = a \leftrightarrow q(y))$.

The main result of this section is the following:

Theorem 11: If the definition of randomness used here is correct, then $\mathbf{M}_0$ is not a possible mathematical universe for quantum mechanics.

Proof: Assume the converse and let s, b be such that in $\mathbf{M}_0$ $0 < p_{sb}(\{1\}) < 1$, Eq. (3). Clearly such s and b exist. By condition (b), $\psi_{tsb} \in M_0$ and $P_{sb} = \otimes\, p_{sb}$ is correct for ψ_{tsb} with p_{sb} nonatomic.

Now each element of M_0 is definable inside $\mathbf{M}_0$.[1,25,26] Thus let q be the formula such that $\mathbf{M}_0 \models \forall x(x = \psi_{tsb} \leftrightarrow q(x))$ which is equivalent to[15] $\forall x \in M_0(x = \psi_{tsb} \leftrightarrow q^{M_0}(x))$, where q^{M_0} is the formula obtained by restricting all quantifiers in q to range over M_0. Let q'_{M_0} be the formula $q'_{M_0}(x) \equiv (x \in M_0 \wedge q^{M_0}(x)) \vee (x \notin M_0 \wedge x \neq x)$. Then $\forall x(x = \psi_{tsb} \leftrightarrow q'_{M_0}(x))$ holds. Now M_0 is also definable.[1,26] Thus let θ be the formula which defines M_0 by $\forall x(x \in M_0 \leftrightarrow \theta(x))$. Let $S(x)$ be the formula obtained from $q'_{M_0}(x)$ by replacing all $\forall z \in M_0$ and $\exists w \in M_0$ by $\forall z\, \theta(z)$ and $\exists w\, \theta(w)$, respectively. Then $\forall x(x = \Psi_{tsb} \leftrightarrow S(x))$ holds and thus ψ_{tsb} is definable. Then $\{0,1\}^\omega - \{\psi_{tsb}\}$ is also definable in $\mathcal{L}_{\mathrm{ZF}}$ (by $x \in \{0,1\}^\omega \vee \neg S(x)$). Since P_{sb} is nonatomic, $P_{sb}(\{0,1\}^\omega - \{\psi_{tsb}\}) = 1$ and, by the correctness requirement, $\psi_{tsb} \in \{0,1\}^\omega - \{\psi_{tsb}\}$, which is impossible. Thus $\mathbf{M}_0$ is not a possible universe for quantum mechanics. QED

Examination of the proof of this theorem shows that in essence what is proved is that if P is a nonatomic product measure and ψ is random for P, then $\psi \notin M_0$. It follows from this that $\mathbf{M}_0$ is not acceptable for only those s and b for which $0 < \mathrm{Tr}_{M_0}(\Psi_{M_0}(s)(\Phi_{M_0}(b)) < 1$. If s and b are such that $\mathrm{Tr}_{M_0}(\Psi_{M_0}(s)(\Phi_{M_0}(b)) = 0$ or $= 1$, then ψ_{tsb} is a constant sequence of 0's or a constant sequence

of 1's respectively and $\psi_{tsb} \in M_0$. For these measurements, M_0 is an acceptable universe for the description of the mathematics of the measurement and its infinite repetition.

One consequence of this is that there are differences between classical and quantum mechanics with respect to the suitability of $\mathbf{M}_0$. First one notes that conditions (a) and (b) are also necessary conditions which classical mechanics must satisfy provided one replaces $B(H)$ in condition (a) by, say, the C^*-algebra of bounded continuous real valued functions over phase space.

It follows from the above that in classical mechanics for any $s \in \mathrm{Dom}\Psi_{M_0}$ such that $\Psi_{M_0}(s)$ is a pure state, M_0 is suitable for *every* $b \in \mathrm{Dom}\Phi_{M_0}$. The reason is that the range set of Φ_{M_0} is a Boolean algebra of questions and, if $\Psi_{M_0}(s)$ is pure, for every $b \in \mathrm{Dom}\Phi_{M_0}$ and every infinite repetition (tsb) of s and b, ψ_{tsb} is either a constant 0 sequence or a constant 1 sequence.

This is not true in quantum mechanics. In this case, if $\Psi_{M_0}(s)$ is pure, then $\mathbf{M}_0$ is not suitable for every $b \in \mathrm{Dom}\Phi_{M_0}$. It is suitable only for those b for which $\Psi_{M_0}(s)$ is dispersion free for $\Phi_{M_0}(b)$.

V. DISCUSSION

It is important to recognize that the proof that $\mathbf{M}_0$ is not a suitable mathematical universe for quantum mechanics depends on how one defines randomness. This is expressed explicitly in Theorem 11 as an if—then statement.

For definitions of randomness stronger than the one used here, it is clear that the conclusions of this paper will still hold. The reason is that any sequence random under a stronger definition is random under the definition used here. However, definitions which are either weaker than or not comparable with the one used here, and which cannot be rejected on some other grounds, must be examined individually.

For example, consider the following definition which is a slight generalization of one used by Solovay.[23] A sequence ψ is M_0-random if there exists a nonatomic M_0 product measure P such that, for all Borel subsets B of $\{0,1\}^\omega$, if B has a code in M_0 and $PB=1$, then $\psi \in B$. $P=\otimes p$ is a nonatomic M_0 measure if $0<p(\{1\})<1$ and $p \in M_0$ [for a description of the coding see Solovay[23] or Jech[24]].

This definition is weaker than the one used here.[27] However, it is easy to show that[28] if ψ is M_0 random, then $\psi \notin M_0$. Thus if this definition of randomness is correct, $\mathbf{M}_0$ is still not a possible mathematical universe for quantum mechanics.

However, there are still weaker definitions such as those given by Martin Lof,[29] which avoid the difficulties[30] of definitions based on subsequence selection procedures and are such that one can show that sequences which are random in this sense exist in M_0. For these definitions the proof given in Theorem 11 fails, and it does not seem possible to exclude $\mathbf{M}_0$ as a possible universe for quantum mechanics by use of arguments based on randomness.

An important difference between these weaker definitions and the stronger one used here is that the latter includes "esoteric" properties which a random sequence should have. For example, the following property is included in the definition used here. There is a well ordering of the elements of the constructible universe L, Eq. (12), which is definable in $\mathcal{L}_{\mathrm{ZF}}$. Since $\mathbf{M}_0 \models$ (every element is constructible)[1,25] and constructibility is absolute,[4] every element of M_0 is constructible outside M_0.

Let ψ_0 be the first constructible 0—1 sequence in the definable well ordering which is not in M_0. Such a ψ_0 exists since: (1) inside $\mathbf{L}$, M_0 is countable[25,26] and $\{0,1\}^\omega$ is uncountable; (2) "ψ_0 is constructible" is absolute. Thus ψ_0 is definable in $\mathcal{L}_{\mathrm{ZF}}$. Since $\{0,1\}^\omega - \{\psi_0\}$ is a set of measure 1 for every nonatomic product measure, the definition of randomness used here excludes ψ_0 whereas ψ_0 is not excluded in the definitions of Martin Lof.

These considerations stress the importance of determining which definition of randomness is correct. On philosophical grounds, the author prefers strong definitions of the sort used here since a random sequence of outcomes should not be definable even by an "esoteric" definition as in the example above. However, at present one cannot reject intermediate definitions such as those of Martin Lof. What is desired is a proof that the correct definition of randomness is at least as strong as (—). That is, that there exists a weakest possible definition. It is speculated here that such a proof will not be forthcoming until one develops a theory to treat both physics and mathematics together as a coherent unit rather than as two separate disciplines.

The proof that M_0 is not a possible mathematical universe is from a mathematical logical viewpoint, nothing new. There is no suggestion in the literature that the real universe, V, of sets, if such exists, is as small and simple as M_0. However, the point that can be made is that if there exists a universe V of sets as a ZFC model, which is more real than the others, then why V is one ZFC model and not another needs explaining. This work suggests that physics may have something to say about this problem.

ACKNOWLEDGMENT

The author wishes to thank Robert Solovay for discussions about various concepts in mathematical logic. In particular discussions about the relative strengths of definitions of randomness as well as pointing out the existence of Refs. 18 and 25 were very helpful.

*Work performed under the auspices of the U.S. Energy Research and Development Administration.

[1]Paul J. Cohen, *Set Theory and the Continuum Hypothesis* (Benjamin, New York, 1966).
[2]P.A. Benioff, Phys. Rev. D 7, 3603 (1973); J. Math. Phys. 15, 552 (1974).
[3]J. Shoenfield, *Mathematical Logic* (Addison-Wesley, Reading, Mass., 1967).
[4]T. J. Jech, *Lectures on Set Theory*, Lecture Notes in Mathematics No. 217 (Springer-Verlag, Berlin, 1971).
[5]Gaisi Takeuti and Wilson M. Zaring, *Introduction to Axiomatic Set Theory* (Springer-Verlag, Berlin, 1971).

[6]Gaisi Takeuti and Wilson M. Zaring, *Axiomatic Set Theory* (Springer-Verlag, Berlin, 1973).
[7]Ulrich Felgner, *Models of ZF Set Theory*, Lecture Notes in Mathematic, No. 223 (Springer-Verlag, Berlin, 1971).
[8]A. Mostowski, *Constructible Sets with Applications* (North-Holland, Amsterdam and Polish Scientific Publishers, Warsaw, 1969).
[9]One can also start the collection process with urelements or atoms. Such set theories with atoms (Ref. 4) will not be pursued further here.
[10]See Ref. 4, pp. 19, 20.
[11]J. Kelley, *General Topology* (Van Nostrand, New York, 1955), Appendix.
[12]A. Frankel, Y. Bar Hillel, and A. Levy, *Foundations of Set Theory* (North-Holland, Amsterdam, 1973), 2nd ed., Chap. 7.
[13]See Ref. 4, pp. 20—21 and Ref. 5, Chap. 12.
[14]See Ref. 6, Chap. 7.
[15]See Ref. 4, pp. 20—24 and Ref. 5, Chap. 13.
[16]Some recent work in physics using nonstandard real numbers [Peter Keleman and Abraham Robinson, J. Math. Phys. 13, 1870, 1875 (1972); A. Voros, J. Math. Phys. 14, 292 (1973); Ryouichi Kambe, Progr. Theoret. Phys. 52, 688 (1974)] has taken explicit recognition of the fact that a similar situation holds for the real numbers axiomatized as a complete ordered field.
[17]C. Chang and H. Keisler, *Model Theory* (North-Holland, Amsterdam, 1973), Chap. I and Appendix A.
[18]L. Pontriagin, *Topological Groups*, translated by Arlen Brown (Gordon and Breach, New York, 1966), 2nd ed., Chap. 14.
[19]The Lowenheim Skolem Tarski theorems do not apply here as topolical concepts are not first order axiomatizable. The author is indebted to Robert Solovay for discussions on this point and for Ref. 18.
[20]H. Ekstein, Phys. Rev. 153, 1937 (1967); 184, 1315 (1969); Y. Avishai and H. Ekstein, Commun. Math. Phys. 37, 193 (1974).
[21]K. Yosida, *Functional Analysis* (Springer-Verlag, Berlin, 1965), Chap. I, Sec. 10.
[22]P. Halmos, *Measure Theory* (Van Nostrand, Princeton, N.J., 1950), Sec. 38.
[23]R. Solovay, Ann. Math. 92, 1 (1970).
[24]See Ref. 4, pp. 80, 81.
[25]Y. Suzuki and G. Wilmers, *Non-Standard Models for Set Theory*, Proceedings of the Bertrand Russell Memorial Logic Conference, Uldum, Denmark, 1971, edited by John Bell, Julian Cole, Graham Priest, and Alan Slomsen, pp. 278—314. The author is indebted to Robert Solovay for this reference.
[26]Robert Solovay, private communication.
[27]One must prove that every Borel set with a code in M_0 is definable. By the proof of Theorem 11 every code $f \in M_0$ is definable. By Ref. 23, the Borel set coded by f is definable. The author is indebted to Robert Solovay for pointing out the relation between the two definitions of randomness.
[28]Assume ψ is M_0 random, and $\psi \in M_0$. Then $\{\psi\}$ has a code in M_0 and thus so does $\{0,1\}^\omega - \{\psi\}$. Let P be the M_0 product measure which satisfies the definition of M_0 randomness. Since P is nonatomic, $P(\{0,1\}^\omega - \{\psi\}) = 1$ which gives $\psi \in \{0,1\}^\omega - \{\psi\}$, a contradiction.
[29]P. Martin Lof, *On the Motion of Randomness*, Proceedings of Summer Institute on Proof Theory and Intuitionism, State University of New York, Buffalo, 1968, edited by J. Myhill, A. Kino and R. Vesley (North-Holland, Amsterdam, 1970); Inform. Control 9, 603 (1966).
[30]J. Ville, *Etude critique de la notion de collectif* (Gauthier-Villars, Paris, 1939).

Models of Zermelo Frankel set theory as carriers for the mathematics of physics. II*

Paul A. Benioff

Chemistry Division, Argonne National Laboratory, Argonne, Illinois 60439
(Received 3 July 1975)

This paper continues the study of the use of different models of ZF set theory as carriers for the mathematics of quantum mechanics. The basic tool used here is the construction of Cohen extensions of ZFC models by use of Boolean valued ZFC models [C = axiom of choice]. Let $\mathbf{M}$ be a standard transitive ZFC model. Inside $\mathbf{M}$, $B(H_M)$ is the algebra of all bounded linear operators over some Hilbert space H_M. It is shown that with each state ρ in $B(H_M)$ and projection operator o in $B(H_M)$ one can associate a unique Boolean valued ZFC model $\mathbf{M}^{\mathcal{B}_{\rho o}}$. $\mathcal{B}_{\rho o}$ is the algebra of all Borel subsets of $\{0,1\}^\omega$, the set of all infinite 0–1 sequences, modulo sets of $P_{\rho o} = \otimes p_{\rho o}$ measure zero with $p_{\rho o}(\{1\}) = \mathrm{Tr}\rho o$ in $\mathbf{M}$. Let Ψ_M and Φ_M be respective maps from the sets of state preparation and question measuring procedures into $B(H_M)$. Let $\mathbf{M} = \mathbf{M}_0$, the minimal standard transitive ZFC model. It is then shown that with each state preparation procedure $s \in \mathrm{Dom}(\Psi_{M_0})$ and each question measuring procedure $q \in \mathrm{Dom}(\Phi_{M_0})$ and with each infinite repetition (tsq) of doing s and q at times $t(0)$, $t(1)$,..., if the definition of randomness is sufficiently strong, one can associate the Cohen extension $\mathbf{M}_0[\psi_{tsq}]$ of $\mathbf{M}_0$ by ψ_{tsq}. ψ_{tsq} is the random outcome sequence associated with (tsq). A third condition, in addition to the two given in the previous paper, is then given which must be satisfied if a ZFC model $\mathbf{M}$ is to serve as a carrier for the mathematics of quantum mechanics. In essence it says that for each pair (tsq) and (wuk) of distinct infinite repetitions of doing s and q and of doing u and k with s, $u \in \mathrm{Dom}(\Psi_M)$ and q, $k \in \mathrm{Dom}(\Phi_M)$, the two outcome sequences ψ_{tsq} and ψ_{wuk} are mutually statistically independent. It is then shown that for a strong definition of independence, corresponding to the definition of randomness used previously, no Cohen extension $\mathbf{M}_0[\psi_{tsq}]$ of $\mathbf{M}_0$ can serve as the carrier for the mathematics of quantum mechanics.

I. INTRODUCTION

In the first paper,[1] hereafter referred to as I, some aspects were discussed of the use of different models of Zermelo Frankel set theory as carriers for the mathematics of quantum mechanics. Among other things it was seen that for strong definitions of random outcome sequences, the minimal ZFC [Zermelo Frankel set of theory with the axiom of choice] model, M_0, cannot serve as a carrier for the mathematics of quantum mechanics.

This paper extends the work of I. The basic tool used here is the construction of Cohen extensions[2–6] of ZFC models by use of the Scott Solovay technique of constructing Boolean valued ZFC models.[4,5,7,8] Section II reviews the main results of I. Section III gives some background material necessary for the remainder of this paper. Part A reviews some of the properties of Boolean valued models of ZFC. In particular the construction of the Boolean valued model $M^{\mathcal{B}}$ from a standard transitive ZFC model $\mathbf{M}$ and an $\mathbf{M}$ complete Boolean algebra $\mathcal{B}$ in M is given along with various properties of relations inside M.

Part B outlines the construction of a Cohen extension $M[G]$ of M from $M^{\mathcal{B}}$ where G, as a subset of $\mathcal{B}$, is an M-generic ultrafilter on $\mathcal{B}$. Part C specializes the general construction of Parts A and B to Boolean algebras which are measure algebras over Borel subsets of $\{0,1\}^\omega_M$ the set of all infinite 0—1 sequences inside $\mathbf{M}$. The main results of this part are given as theorems and lemmas since they are simple generalizations of the results of Solovay[9] for Lebesgue measure on the real line, to arbitrary probability measures on $\{0,1\}^\omega_M$. The main results here (Theorem 5) says that to each P such that, inside $\mathbf{M}$, P is a probability measure on $\mathcal{B}(\{0,1\}^\omega_M)$ there corresponds, outside $\mathbf{M}$, a unique probability measure Q on $\mathcal{B}(\{0,1\}^\omega)$. In Part D the association between sequences for which some measure Q is correct and M-generic ultrafilters on $\mathcal{B}_P$ is given. It is shown (Theorem 7) that if $\mathbf{M}$ is countable and P and Q are related by Theorem 5, then there is a canonical one—one correspondence between the set of sequences in $\{0,1\}^\omega$ for which Q is M correct and the set of M-generic ultrafilters over $\mathcal{B}_P$.

The results of Sec. III are applied to quantum mechanics in Sec. IV. In Part A quantum mechanics based on $\mathbf{M}$ with $\mathbf{M}$ an arbitrary standard transitive ZFC model is considered. The main result, Theorem 11, says the following: With each state ρ and question observable o in $B(H_M)$ where inside $\mathbf{M}$, $B(H_M)$ is the set of all bounded linear operators over some Hilbert space H_M, there is associated a unique Boolean valued ZFC model $\mathbf{M}^{\mathcal{B}_{\rho o}}$ where $\mathcal{B}_{\rho o}$ is the measure algebra over $\mathcal{B}(\{0,1\}^\omega_M)$ constructed inside $\mathbf{M}$ from the product measure $P_{\rho o} = \otimes p_{\rho o}$, where $p_{\rho o}(\{1\}) = \mathrm{Tr}_M(\rho o)$. This theorem is discussed. Several ways of generalizing it, which may also be relevant for quantum mechanics, are discussed.

In I two necessary conditions were given which must be satisfied if $\mathbf{M}$ is to serve as a carrier for the mathematics of quantum mechanics: (a) Every state preparation procedure s and every question measuring procedure q corresponds, respectively, to a density operator $\psi_M(s)$ and a projection operator $\Phi_M(q)$ in $B(H_M)$; (b) With each (tsq) such that (tsq) is an infinite repetition of carrying out s and q at time $t(0), t(1), \cdots$ there is associated a random outcome sequence $\psi_{tsq} \in M$ such that inside $\mathbf{M}$, $\bar{M}\psi_{tsq} = \mathrm{Tr}_M(\psi_M(s)\Phi_M(q))$.

In Part B $\mathbf{M}$ is restricted to be $\mathbf{M}_0$, the minimal standard transitive ZFC model. Conditions (a) and (b) are used with the previous results to prove that (Theorem 12) with each $s \in \mathrm{Dom}(\psi_{M_0})$ and $q \in \mathrm{Dom}(\Phi_{M_0})$ such that

$0 < \mathrm{Tr}_{M_0}(\psi_{M_0}(s)\Phi_{M_0}(q)) < 1$ and with each $t \in M_0$ such that (tsq) is an infinite repetition of doing s and q at $t(0), \cdots$, for a sufficiently strong definition of randomness, there can be associated the unique standard transitive ZFC model $M_0[\psi_{tsq}]$, where $M_0[\psi_{tsq}]$ is the Cohen extension of $\mathbf{M}_0$ by ψ_{tsq}.

In Sec. V another condition which a model **M** must satisfy is given. It is that for any state preparation procedures s and $u \in \mathrm{Dom}(\Psi_M)$ with $s \neq u$ and any pair of question measuring procedures q and $k \in \mathrm{Dom}(\Phi_M)$ with $q \neq k$ such that $0 < \mathrm{Tr}_M(\Psi_M(s)\Phi_M(q)) < 1$ and $0 < \mathrm{Tr}_M(\Psi_M(u)\Phi_M(k)) < 1$, and for all t and $w \in M$ such that (tsq) and (wuk) are respectively infinite repetitions of doing s and q and doing u and k, ψ_{tsq} and ψ_{wuk} are mutually statistically independent. It is proved that (Theorem 15) if a strong definition of statistical independence is correct then no model of the type $M_0[\psi_{tsq}]$ can serve as the carrier of the mathematics of quantum mechanics.

In Sec. VI these results are discussed. Among other things it is noted that an important open problem is to determine which definition of randomness and independence is correct.

II. REVIEW OF I

In I, some aspects of a ZFC model as a carrier for the mathematics of physics as specialized to quantum mechanics were discussed. The following two conditions were given as necessary (but not sufficient) for a standard transitive ZFC model **M** to serve as the carrier for the mathematics of quantum mechanics.

(a) Let $\mathcal{S}$ and $\mathcal{Q}$ denote the respective sets of state preparation procedures and question measuring procedures. Then there exist maps $\Psi_M : \subseteq \mathcal{S} \to B(\mathcal{H}_M)$ and $\Phi_M : \subseteq \mathcal{Q} \to B(\mathcal{H}_M)$ such that for each $s \in \mathrm{Dom}(\Psi_M) \subseteq \mathcal{S}$ and each $q \in \mathrm{Dom}(\Phi_M) \subseteq \mathcal{Q}$, $M \models \Psi_M(s)$ and $\Phi_M(q)$ are respective density operators and projection operators in $B(\mathcal{H}_M)$. [$\mathbf{M} \models \mathcal{Q}$ means the formula $\mathcal{Q}$ is true inside M.]

(b) Inside **M**, let R_M be the set of reals and let $t : \omega \to R_M$ be an increasing function. Then for each $s \in \mathrm{Dom}(\Psi_M)$, $q \in \mathrm{Dom}(\Phi_M)$ and $t \in M$ such that (tsq) is an infinite repetition of doing s followed by q [where for each j the jth repetition is done at time $t(j)$], there exists a sequence $\psi_{tsq} \in \{0, 1\}^\omega$ such that: (1) $\psi_{tsq} \in M$ and ψ_{tsq} is random. (2) Inside **M**

$$\bar{M}\psi_{tsq} = \mathrm{Tr}_M(\Psi_M(s)\Phi_M(q)), \tag{1}$$

where $\bar{M}$ denotes the limit mean and Tr_M is the trace operation on $B(\mathcal{H}_M)$ in **M**. (3) For each m the outcome sequence obtained by doing the first m repetitions of (tsq) is given by the first m elements of ψ_{tsq}.

These conditions were discussed. In particular let QM_M and QM denote, respectively, quantum mechanics based in **M** and quantum mechanics based in the usual intuitive mathematics. [Conditions (a) and (b) for QM are obtained by deleting all references to **M** in the above.]

The following theorem was proved:

Theorem 1: Let $\mathcal{H}_M$ and $B(\mathcal{H}_M)$ be as in condition (a). Then there exists outside M a Hilbert space $\mathcal{H}$ and the operator algebra $B(\mathcal{H})$ and maps $U_M : \mathcal{H}_M \to \mathcal{H}$ and $V_M : B(\mathcal{H}_M) \to B(\mathcal{H})$ such that U_M and V_M are isometric monomorphisms.

From this theorem a natural correspondence between the maps Ψ_M and Φ_M of QM_M and Ψ and Φ of QM is given by requiring that $\mathrm{Dom}(\Psi_M) \subseteq \mathrm{Dom}(\Psi)$, $\mathrm{Dom}(\Phi_M) \subseteq \mathrm{Dom}(\Psi)$, and

$$\begin{aligned} \Psi(s) &= V_M(\Psi_M(s)), \\ \Phi(q) &= V_M(\Phi_M(q)) \end{aligned} \tag{2}$$

for all $s \in \mathrm{Dom}(\Psi_M)$ and $q \in \mathrm{Dom}(\Phi_M)$. It follows that

$$\mathrm{Tr}(\Psi(s)\Phi(q)) = \mathrm{Tr}_M(\Psi_M(s)\Phi_M(q)) \tag{3}$$

holds for all $s \in \mathrm{Dom}(\Psi_M)$ and $q \in \mathrm{Dom}(\Phi_M)$.

It was noted that Theorem 1 and Eqs. (2) and (3) also hold between any two standard transitive ZFC models **M** and **N** with $M \subseteq N$ and not just between M and the real world **V** of ZF sets. That is, to $\mathcal{H}_M$ and $B(\mathcal{H}_M)$ inside **M** there corresponds a $\mathcal{H}_N$ and $B(\mathcal{H}_N)$, such that inside N, $\mathcal{H}_N$ is a Hilbert space and $B(\mathcal{H}_N)$ is the algebra of all bounded linear operators over $\mathcal{H}_N$, along with isometric monomorphisms $U_{NM} : \mathcal{H}_M \to \mathcal{H}_N$ and $V_{NM} : \mathcal{H}_M \to \mathcal{H}_N$. Similarly for QM_M and QM_N, $\mathrm{Dom}(\Psi_M) \subseteq \mathrm{Dom}(\Psi_N)$ and $\mathrm{Dom}(\Phi_M) \subseteq \mathrm{Dom}(\Phi_N)$. Also

$$\begin{aligned} \Psi_N(s) &= V_{NM}\Psi_M(s), \\ \Phi_N(q) &= V_{NM}\Phi_M(s) \end{aligned} \tag{2'}$$

and

$$\mathrm{Tr}_N(\Psi_N(s)\Phi_N(q)) = \mathrm{Tr}_M(\Psi_M(s)\Phi_M(q)) \tag{3'}$$

for all $s \in \mathrm{Dom}(\Psi_M)$ and $q \in \mathrm{Dom}(\Phi_M)$.

Theorem 1 is used to construct the correspondences of Eqs. (2) and (3) only in the absence of any requirements on $\mathcal{H}_M$ and $B(\mathcal{H}_M)$ other than those given in condition (a). If there are further requirements, for example that $\mathbf{M} \models \mathcal{H}_M = L^2(R, \beta R), \mu)_M$, then outside **M** one requires that $\mathcal{H} = L^2(R, \beta(R), \mu)$. Then the maps U_M and V_M must be changed to correspond to this situation.

All the mathematical results and theorems used so far in quantum mechanics (and in physics) can be cast as results and theorems of ZFC. Since they are true in every ZFC model, all ZFC models should be equivalent as carriers for the mathematics of quantum mechanics. This includes the real world **V** as well as any other ZFC model **M**.[10]

It was then shown that this may not be correct. In particular the following theorem was proved.

Theorem 2: If a strong definition of randomness is correct, then the minimal standard ZFC model, $\mathbf{M}_0$, cannot serve as the carrier for the mathematics of physics.

In particular the theorem was proved, using condition (b), for the following strong definition,[11,12] of randomness. A sequence ψ is *random* if there exists a product measure $P = \otimes p$ on $\beta(\{0, 1\}^\omega)$, the set of Borel subsets of $\{0, 1\}^\omega$, with $p \in M_0$ such that P is correct for ψ. P is *correct* for ψ if for all Borel subsets B of $\{0, 1\}^\omega$, if B is definable from P in $\mathcal{L}_{ZF}$, the language of set theory and $PB = 1$, then $\psi \in B$.

It was noted that the proof of the theorem also holds for the definition of randomness which is in essence that given by Solovay[9] and fails for the weaker definitions given by Martin Löf.[13] An important open question is to discover which definition of randomness is correct.

III. COHEN EXTENSIONS THROUGH BOOLEAN VALUED ZFC MODELS

A. Boolean valued models of ZFC

Here some properties of Boolean valued models of ZF set theory are briefly reviewed. For a more complete treatment the reader is referred to the literature.[4,5,7,8]

Let M be a standard transitive model of ZFC. Let $\mathcal{B} \in \mathbf{M}$ be a Boolean algebra which is complete in M; that is, for all $A \subset \mathcal{B}$ and $A \in M$, Inf A and Sup A exist in $\mathcal{B}$. [That is, inside M, $\mathcal{B}$ is a complete Boolean algebra.] Define $M^{\mathcal{B}}$ as follows.[4,5] For each ordinal α

$$M_0^{\mathcal{B}} = 0 \quad \text{(the empty set)}, \tag{4a}$$

$$M_{\alpha+1}^{\mathcal{B}} = \{u \mid u \in M \text{ and } u \text{ is a function with domain } \subseteq M_\alpha^{\mathcal{B}} \text{ and range } \subseteq \mathcal{B}\}, \tag{4b}$$

$$M_\alpha^{\mathcal{B}} = \bigcup_{\beta<\alpha} M_\beta^{\mathcal{B}} \quad \text{if } \alpha \text{ is a limit ordinal}, \tag{4c}$$

$$M^{\mathcal{B}} = \bigcup_{\alpha \in On_M} M_\alpha^{\mathcal{B}},$$

where On_M is the class of all ordinals in M. $M^{\mathcal{B}}$ is the class of all $\mathcal{B}$ valued functions in M which are predicatively defined. That is, each $u \in M^{\mathcal{B}}$ is defined in terms of previously defined functions only.

Let Q denote any formula in the language of set theory and let $\|Q\|$ denote the Boolean value of Q in $M^{\mathcal{B}}$. That is, for each Q, $\|Q\| \in \mathcal{B}$. $\|Q\|$ is defined as follows.[4,5] For each $u \in M^{\mathcal{B}}$ let the rank, $\rho(u)$, of u be the least ordinal α such that $u \in M_{\alpha+1}^{\mathcal{B}}$. Let u and v be elements of $M^{\mathcal{B}}$. One defines $\|u=v\|$ and $\|u \in v\|$ by simultaneous recursion on $\rho(u) \times \rho(v)$ as follows:

$$\|u \in v\| = \sum_{\omega \in \mathrm{Dom}\, v} v(\omega) \cdot \|\omega = u\|, \tag{5a}$$

$$\|u=v\| = \prod_{\omega \in \mathrm{Dom}(u)} (u(\omega) \Rightarrow \|\omega \in v\|) \times \prod_{z \in \mathrm{Dom}(v)} ((z) \Rightarrow \|z \in u\|) \ . \tag{5b}$$

In these expressions $\sum$ and $\prod$ denote respectively the least upper bound and the greatest lower bound in $\mathcal{B}$. $\mathrm{Dom}(u)$ = domain of u, $\cdot$ denotes the Boolean "and," and $b \Rightarrow c \equiv b^\perp + c$, with $+$ the Boolean "or" and $\perp$ the Boolean negation, denotes the Boolean "implies." Expressions (5) are the Boolean equivalents of $u \in v \leftrightarrow \exists w (w \in v \wedge w = u)$ and $u = v \leftrightarrow \forall w (w \in u \leftrightarrow w \in v) \wedge \forall z (z \in v \leftrightarrow z \in u)$ of ordinary two valued logic.

For any formulas Q and Q' of $\mathcal{L}_{ZF}$ one has the following[4,5]:

$$\|\neg Q\| = \|Q\|^\perp, \tag{6a}$$

$$\|Q \wedge Q'\| = \|Q\| \cdot \|Q'\|, \tag{6b}$$

$$\|Q \vee Q'\| = \|Q\| + \|Q'\|, \tag{6c}$$

$$\|Q \rightarrow Q'\| = \|Q\| \Rightarrow \|Q'\|, \tag{6d}$$

$$\|\exists u Q\| = \sum_u \|Q(u)\|, \tag{6e}$$

$$\|\forall u Q\| = \prod_u \|Q(u)\|, \tag{6f}$$

where $\sum$ and $\prod$ are over all $u \in M$.

From these definitions, many properties can be obtained,[4,5] and a few are given here. One has $u \in \mathrm{Dom}(v) \rightarrow v(u) \leq \|u \in v\|$, $\|u=u\| = \mathbf{1}$ the unit of $\mathcal{B}$, $\|u=v\| \cdot Q(u)\| \leq \|Q(v)\|$, $\|\exists u \in v (Q(u))\| = \sum_{u \in \mathrm{Dom}(v)} v(u) \cdot \|Q(u)\|$, and $\|\forall u \in v (Q(u))\| = \prod_{u \in \mathrm{Dom}(v)} v(u) \Rightarrow \|Q(u)\|$ for all $u, v \in M^{\mathcal{B}}$. One has the maximum principle[8,4] which says that if a formula is of the form $\exists x Q(x)$ then there exists an element $v \in M^{\mathcal{B}}$ such that $\|Q(v)\| = \|\exists u Q(u)\|$, v is unique if $M^{\mathcal{B}}$ is separated (see below).

An element $v \in M^{\mathcal{B}}$ is extensional if for all $u \in \mathrm{Dom}(v)$, $v(u) = \|u \in v\|$.[4] Note that in general $u \notin \mathrm{Dom}(v)$ does not imply $\|u \in v\| = 0$. Also one can have $\|u=v\| = \mathbf{1}$ with $u \neq v$. A Boolean valued structure $M^{\mathcal{B}}$ for which $\|u=v\| = \mathbf{1}$ implies $u=v$ is called separated. One can always construct a separated structure $M_s^{\mathcal{B}}$ from $M^{\mathcal{B}}$ by letting $M_s^{\mathcal{B}} = \{[v] \mid v \in M\}$ with $[v] = \{u \mid \|u=v\| = \mathbf{1}$ and u is of minimal rank$\}$. $\in$ and $=$ are lifted onto $M_s^{\mathcal{B}}$ by defining $\|[u] \in [v]\| = \|u \in v\|$ and $\|[u] = [v]\| = \|u=v\|$ for all $u, v \in M^{\mathcal{B}}$. For each formula Q one has then $\|Q(u_1 \cdots u_n)\| = \|Q([u_1] \cdots [u_n])\|$.

Let $\mathbf{M}^{\mathcal{B}} = \langle M^{\mathcal{B}}, \in, = \rangle$ with $\in$, $=$ defined as in Eq. (5). One says that a formula Q is "true" in $\mathbf{M}^{\mathcal{B}}$ or "$\mathcal{B}$ valid" in $\mathbf{M}^{\mathcal{B}}$ if $\|Q\| = \mathbf{1}$. As in the 2 valued case this is denoted by $\mathbf{M}^{\mathcal{B}} \models Q$. $\mathbf{M}^{\mathcal{B}}$ is a $\mathcal{B}$ valued model of ZFC. That is all axioms (and theorems) of ZFC are $\mathcal{B}$ valid in $\mathbf{M}^{\mathcal{B}}$.

There is an embedding of M into $M^{\mathcal{B}}$ defined as follows.[4,5] Let M^2 be defined by Eq. (4) with the two element Boolean algebra $\mathbf{2} = \{0, 1\}$ replacing $\mathcal{B}$. M^2 is the class of all functions in M which have values in 2. Let $\check{M}$ be the class of all functions in M which have value 1 only. One clearly has $\check{M} \subset M^2 \subset M$. There is a natural isomorphism between M and $\check{M}$ given by the $\in$ recursion,[4,5]

$$\check{0} = 0 \quad \text{(the empty set in } M), \tag{7a}$$

$$\mathrm{Dom}(\check{x}) = \{\check{y} \mid y \in x\},$$

$$\check{x}(\check{y}) = \mathbf{1} \quad \text{for all } \check{y} \in \mathrm{Dom}(\check{x}). \tag{7b}$$

One has $\|\check{x} = \check{y}\| = \mathbf{1} \leftrightarrow x = y$, $\|\check{x} = \check{y}\| = \mathbf{0} \leftrightarrow x \neq y$, $\|\check{x} \in \check{y}\| = \mathbf{1} \leftrightarrow x \in y$, and $\|\check{x} \in \check{y}\| = \mathbf{0} \leftrightarrow x \in y$ for all $x, y \in M$.

Inside M the (standard) natural numbers are $\check{0}, \check{1}, \cdots, \check{n}, \cdots$, where $0, 1, \cdots, n$ are the natural numbers. That is, let $Q(x)$ be the ZF formula which says "x is a natural number." Then $\|Q(u)\| = \sum_{n \in \omega} \|\check{n} = u\|$ and $\|Q(\check{n})\| = \mathbf{1}$ for each $n \in \omega$. Similarly for each real number r in M one has $\check{r}$ in $M^{\mathcal{B}}$. Similar constructions hold for other objects such as 0–1 sequences, etc.

$M^{\mathcal{B}}$ also contains many nonstandard objects.[14] For example,[4] let $b^\perp$ be the complement of some element $b \in \mathcal{B}$ and define v by $v = b \cdot \check{m} + b^\perp \cdot \check{n}$ with $m \neq n$. That is, $v(\check{k}) = b \cdot \check{m}(\check{k}) + b^\perp \cdot \check{n}(\check{k})$ for all $k < \max(n, m)$. Inside $\mathbf{M}^{\mathcal{B}}$ v is a natural number as $\|Q(v)\| = \mathbf{1}$. But v corresponds to neither n nor m of M.

Inside $\mathbf{M}^{\mathcal{B}}$ one can express "$u \in M$" by a formula $Q_M(u)$ whose[4,5] Boolean value is given by $\|Q_M(u)\| = \sum_{x \in M} \|\check{x} = u\|$. The sum in this can be limited to range over all x for which $\check{x} \in M^{\mathcal{B}}_{\rho(u)}$, Eq. (4), i.e., all x with rank $\leq$ rank u. Note too that for the v defined above $\|Q_M(v)\| = 1$ or $M^{\mathcal{B}} \models v \in M$ holds. If $\mathcal{B}$ is nonatomic [which is the main type of interest here], then $\mathbf{M}^{\mathcal{B}}$ also contains many elements which, *inside* $\mathbf{M}\mathcal{B}$, are not elements of M. For example, let $\mathcal{G}$ be the function defined by $\mathrm{Dom}(\mathcal{G}) = \{\check{b} \mid b \in \mathcal{B}\}$ and $\mathcal{G}(\check{b}) = b$ for each $b \in \mathcal{B}$. Then one has[4,5] $\|Q_M(\mathcal{G})\| = 0$ if and only if $\mathcal{B}$ is nonatomic.

Another example which will be referred to later is defined as follows: For each $n \in \omega$ and $j \in \{0, 1\}$ let $\langle \check{n}, \check{j} \rangle^{\check{}}$ denote the function whose domain is $\langle \check{n}, \check{j} \rangle = \{\{\check{n}\}^{\check{}}, \{\check{n}, \check{j}\}^{\check{}}\}$, and whose range set is $\{1\}$. That is, "$\langle \check{n}, \check{j} \rangle^{\check{}}$ is an ordered pair in M of natural numbers n and $j \in \{0, 1\}$" is $\mathcal{B}$ valid. Define ψ as follows:

$$\mathrm{Dom}(\psi) = \{\langle \check{n}, \check{j} \rangle^{\check{}} \mid n \in \omega,\ j \in \{0, 1\}\}, \tag{8a}$$

$$\psi(\langle \check{n}, \check{0} \rangle^{\check{}}) = (\psi(\langle \check{n}, \check{1} \rangle^{\check{}}))^{\perp}. \tag{8b}$$

Then one can show that $\|\psi$ is an infinite 0—1 sequence$\| = 1$ and $\|Q_M(\psi)\| = 0$ if and only if

$$\prod_{n \in \omega} \psi(\langle \check{n}, \phi(n) \rangle^{\check{}}) = 0 \tag{9}$$

for each $n \in \omega$ and for each $\phi \in \{0, 1\}^{\omega}_{M}$. {*Proof*: $\|\psi$ is an infinite 0—1 sequence$\| = 1$ follows from $\|\langle \check{n}, \check{0} \rangle^{\check{}} \in \psi \vee \langle \check{n}, \check{1} \rangle^{\check{}} \in \psi\| = 1$ for each n and $\|\langle \check{n}, \check{0} \rangle^{\check{}} \in \psi \wedge \langle \check{n}, \check{1} \rangle^{\check{}} \in \psi\| = 0$ for each n. Next $\|Q_M(\psi)\| = \sum_{x \in M} \|\check{x} = \psi\|$. By Eq. (5b) and the definition of Q_M and $\check{x}$, $\|\check{x} = \psi\| = \prod_{y \in x} \|\check{y} \in \psi\| \cdot \prod_{\langle n,j \rangle \in \omega \times \{0,1\}} (\psi(\langle \check{n}, \check{j} \rangle^{\check{}})^{\perp} + \check{x}(\langle \check{n}, \check{j} \rangle^{\check{}}))$. By Eq. (5a) $\|\check{y} \in \psi\| = 0$ unless $y \in \omega \times \{0, 1\}$. Thus $\|\check{x} = \psi\| = \prod_{\langle n,j \rangle \in x} \psi(\langle \check{n}, \check{j} \rangle^{\check{}}) \cdot \prod_{\langle n,j \rangle \notin x} \psi(\langle \check{n}, \check{j} \rangle^{\check{}})^{\perp}$. From Eq. (8b) one has $\|\check{x} = \psi\| = 0$ if x such that (1) for some n, $\langle n, 0 \rangle \in x$ and $\langle n, 1 \rangle \in x$, or (2) for some n, $\langle n, 0 \rangle \notin x$ and $\langle n, 1 \rangle \notin x$. Thus, $\|Q_M(\psi)\| = \sum_{\phi \in \{0,1\}^{\omega}_{M}} \|\check{\phi} = \psi\| = 0$ if and only if Eq. (9) holds for each $\phi \in \{0, 1\}^{\omega}_{M}$.}

Thus one can see that there are many elements in $M^{\mathcal{B}}$ which, inside $\mathbf{M}^{\mathcal{B}}$, do not belong to M. Recall that, by the construction of $M^{\mathcal{B}}$, $M^{\mathcal{B}} \subset M$. There is no contradiction here since $M^{\mathcal{B}} \subset M$ is a statement outside $\mathbf{M}^{\mathcal{B}}$ and $M \subset M^{\mathcal{B}}$ [both containments are proper] is a statement inside $M^{\mathcal{B}}$.

$\mathbf{M}^{\mathcal{B}}$ also has the following property. Let π be any automorphism of $\mathcal{B}$. π induces an automorphism Π of M given by[4,5] $\mathrm{Dom}(\Pi u) = \{\Pi v \mid v \in \mathrm{Dom}(u)\}$ and

$$(\Pi u)(\Pi v) = \pi(u(v))$$

for all $v \in \mathrm{Dom}(u)$ and for all $u \in M$.

The construction of $M^{\mathcal{B}}$ from M can be iterated.[8] That is, let $\mathcal{C}$ be such that $\mathbf{M}^{\mathcal{B}} \models \mathcal{C}$ is a complete Boolean algebra. Then inside $\mathbf{M}^{\mathcal{B}}$ one can construct $\mathbf{M}^{\mathcal{B} \times \mathcal{C}}$, which is also a Boolean valued ZFC model. This construction can be iterated into the transfinite.[8]

So far the discussion in this section has been carried on outside M. It can, with minor changes, be carried out entirely inside M.[5] From here on, though, the discussion cannot, in general, be carried out inside M.

B. Cohen extensions

Let M be a standard transitive ZFC model and, inside M, let $\mathcal{B}$ be a complete Boolean algebra. An M-generic ultrafilter G on $\mathcal{B}$ is a nonempty subset of $\mathcal{B} - \{0\}$ such that (1) is $b \in G$, and $c \in \mathcal{B}$ and $b \leq c$, then $c \in G$, (2) if $b \in G$ and $c \in G$, then $b \cdot c \in G$, (3) for each $b \in G$, either b or $b^{\perp} \in G$, and (4) for each subset D of G, if $D \in M$, then $\Pi D \in G$. That is, G is an M-complete ultrafilter on the positive elements of $\mathcal{B}$.

In general G is not in M and if $\mathcal{B}$ is nonatomic in M, G is never in M.[5] Also, if $\mathcal{B}$ is countable, then there always exists an M-generic ultrafilter on $\mathcal{B}$.[4] Thus one can always ensure the existence of M-generic ultrafilters on $\mathcal{B}$ by requiring M to be countable.

The Cohen extension $M[G]$ of M by G is defined as follows[5]: For each ordinal α in M one defines an interpretation $\mathcal{J}_{G\alpha}$ of M_{α} [Eq. (4)] as follows:

$$\mathcal{J}_{G0}(0) = 0, \tag{10a}$$

if $\alpha = \beta + 1$, then for each $u \in M^{\mathcal{B}}_{\alpha}$

$$\mathcal{J}_{G\alpha}(u) = \{\mathcal{J}_{G\beta}(v) \mid v \in \mathrm{Dom}(u) \text{ and } u(v) \in G\}, \tag{10b}$$

and if α is a limit ordinal

$$\mathcal{J}_{G\alpha} = \bigcup_{\beta < \alpha} \mathcal{J}_{G\beta}. \tag{10c}$$

Then $\mathcal{J}_G$ is defined as

$$\mathcal{J}_G = \bigcup_{\alpha \in On_M} \mathcal{J}_{G\alpha}, \tag{10d}$$

where On_M is the class of all ordinals in M. $\mathcal{J}_G$ is called the interpretation of $M^{\mathcal{B}}$ [Eq. (4)] by G.[5] Finally $M[G]$ is defined as the range class of $\mathcal{J}_G$. That is,

$$M[G] = \bigcup_{\alpha \in On_M} \{\mathcal{J}_{G\alpha}(v) \mid v \in M^{\mathcal{B}}_{\alpha}\}. \tag{11}$$

$M[G]$ has some interesting properties and some are given here briefly. For any $x \in M[G]$ let $\mathbf{x}$ denote a name for x in $M^{\mathcal{B}}$. That is, $\mathcal{J}_G(\mathbf{x}) = x$. One has that $M \subseteq M[G]$ [*proof*: $\mathcal{J}_G(\check{x}) = x$ for each $x \in M$] and $G \in M[G]$ [*proof*: $\mathcal{J}_G(\mathcal{G}) = G$, where $\mathcal{G}$ was defined in part A].

A very important property of $M[G]$ is that for any formula Q of $\mathcal{L}_{ZF}$

$$M[G] \models Q(a_1 \cdots a_n) \leftrightarrow \|Q(\mathbf{a}_1 \cdots \mathbf{a}_n)\| \in G \tag{12}$$

holds for each $a_1 \cdots a_n \in M[G]$. From this and the fact that $\mathbf{M}^{\mathcal{B}}$ is a Boolean valued ZFC model, one has the result that $\mathbf{M}[G]$ is a ZFC model. M[G] is also a standard transitive model if M is, and is the smallest model such that $M \subseteq M[G]$ and $G \in M[G]$. $M[G]$ is countable if and only if M is, and has the same ordinals as M. Every cardinal of $\mathbf{M}[G]$ is a cardinal of M. If $\mathcal{B}$ satisfies the countable chain condition in M [i.e., in M every subset of $\mathcal{B}$, of pairwise disjoint elements is at most countable], then M and $M[G]$ have the same cardinals.[4]

C. Specialization

The following is in essence a simple extension of the work of Solovay[9,5] to include different probability measures. Let M be a countable standard transitive ZFC model and inside M let $\{0, 1\}^{\omega}_{M}$, $\mathcal{B}(\{0, 1\}^{\omega}_{M})$, and P be respectively, the set of all infinite 0—1 sequences, all

Borel subsets of $\{0,1\}_M^\omega$ and a probability measure on $\mathcal{B}(\{0,1\}_M^\omega)$. Let $\mathcal{B}$ be the measure algebra[15] $\mathcal{B}_P \equiv \mathcal{B}(\{0,1\}_M^\omega)/\mathcal{J}_P$, where in $\mathbf{M}\mathcal{J}_P$ is the ideal of all Borel sets of P measure 0. It is known[4] that $\mathcal{B}_P$ is a complete Boolean algebra which is nonatomic if and only if P is nonatomic [P is nonatomic if $P\{\phi\}=0$ for each $\psi\in\{0,1\}_M^\omega$ and a Boolean algebra $\mathcal{B}$ is nonatomic if for each $b\in\mathcal{B}-\{0\}$ there is a b' such that $0<b'<b$.]

One can now construct, by Eq. (4), the Boolean valued ZFC model $\mathrm{M}^{\mathcal{B}_P}$ in M. To proceed further, it is necessary to relate Borel subsets of $\{0,1\}_M^\omega$ and probability measures inside M to the corresponding objects outside M. This can be done by a coding of the Borel sets into ω^ω, the set of infinite sequences of natural numbers.[9]

To construct the codes, one first notes that the Borel subsets of $\{0,1\}^\omega$ can be defined as follows by induction on the ordinals[5]:

$$\mathcal{B}_0=\{B_{ni}\,|\,n\in\omega \text{ and } i\in\{0,1\} \text{ and } B_{ni}=\{\phi\,|\,\phi(n)=i\}\}$$

For each odd ordinal α

$$\mathcal{B}_\alpha=\{B\,|\,B=\{0,1\}^\omega-B' \text{ where } B'\in\mathcal{B}_\beta \text{ for some } \beta<\alpha.$$

For each even ordinal α

$$\mathcal{B}_\alpha=\{B\,|\,B=\bigcup_m B_m \text{ where for each } m\in\omega, B_m\in\mathcal{B}_{\beta_m} \text{ and } \beta_m<\alpha\},$$

$$\mathcal{B}(\{0,1\}^\omega)=\bigcup_{\alpha<\omega_1}\mathcal{B}_\alpha,$$

where ω_1 is the first uncountable ordinal.

The codes and their corresponding Borel sets are constructed as follows[5,9]:

$$C_0=\{f\,|\,f\in\omega^\omega, f(0)=0 \text{ and } f(2)\in\{0,1\}\} \tag{13a}$$

and for each $f\in C_0$ the Borel set coded by f is given by

$$B_f=\{\phi\,|\,\phi(f(1))=f(2)\}.$$

Let $T:\omega^\omega\to\omega^\omega$ be defined by $(Tf)(j)=f(j+1)$ for each j. Then, if α is odd,

$$C_\alpha=\{f\,|\,f\in\omega,\ f(0)=1 \text{ and } Tf\in C_\beta \text{ for some } \beta<\alpha\} \tag{13b}$$

and for each $f\in C_\alpha$

$$B_f=\{0,1\}^\omega-B_{Tf}.$$

Let J be a one—one map of $\omega\times\omega$ onto $\omega-\{0\}$ [for example, $J(m,n)=(2m+1)2^n$] and for each f and m let f_{Jm} be the sequence defined by $f_{Jm}(n)=f(J(m,n))$ for each n. Let $\beta:\omega\to\alpha$ be a sequence of ordinals all $<\alpha$. If α is even, define C_α by

$$C_\alpha=\{f\,|\,f\in\omega,\ f(0)=2 \text{ and } f_{Jm}\in C_{\beta_m}\} \tag{13c}$$

and for each $f\in C_\alpha$

$$B_f=\bigcup_m B_{f_{Jm}}.$$

Finally define C by

$$C=\bigcup_{\alpha<\omega_1}C_\alpha \tag{13d}$$

From the above it is clear that the map $f\to B_f$ is a map from C onto $\mathcal{B}(\{0,1\}^\omega)$ which preserves the structure. The map is not 1—1 since a Borel set can be constructed in different ways. Also each f encodes its unique decomposition down to the elements of C_0, the generator set.

The reason that one works with codes rather than the Borel sets is that, unlike the Borel sets, the codes are absolute. For example $f\in\omega^\omega$, $f(0)=0$ and $f(2)=1$ is absolute but $B_f=\{\phi\,|\,\phi(f(1))=1\}$ is a different set inside M than it is outside. In general one has the following: "f is a code" is absolute and the structure of f is absolute. That is, "f is a generator code," "f codes the complement of the set coded by Tf," and "f codes the union of the sets coded by f_{Jm} for $m=0,1,\cdots$" are all absolute.[5,9]

The natural correspondence from the Borel subsets of $\{0,1\}_M^\omega$ inside M into the set of Borel subsets of $\{0,1\}^\omega$ outside M is given by the correspondence

$$B_f^M\to B_f \tag{14}$$

for each code f in M. B_f^M and B_f are the Borel sets coded by f which are in M and outside M respectively. Since "ϕ is an infinite 0—1 sequence" is absolute $B_f^M=B_f\cap\{0,1\}_M^\omega$.

The correspondence from probability measures inside M to those outside M is given by the following theorem. First some preliminary definitions and lemmas are needed. Let α be an even, countable ordinal. A code $f\in C_\alpha$ is *n finite* if for all $m\geq n$ $f_{Jm}=f_{Jn}$. Let f be any code in C_α, α even. Define for each n the function $f^n\in\omega^\omega$ by $f^n(0)=2$, for each $m\leq n$, $f^n_{Jm}=f_{Jm}$, and $f^n_{Jm}=f_{Jn}$ for each $m>n$. f^n is called the *n constant* code of f.

Lemma 3: (1) "f is n finite" is absolute.

(2) $f\in C_\alpha\to f^n\in C_\alpha$ for each n.

(3) "f^n is the n constant code of f" is absolute.

Proof: (1) f is a code is absolute[5,9] as is the definition of J, β [Eq. (13c)], and $f_{Jm}=f_{Jn}$.

(2) Obvious from the definition of f^n and C_α.

(3) Follows from 1 and the fact that the structure of the codes is absolute. QED

Lemma 4: Inside M let S be a subset of the real line such that $\sup^M S$ exists. Then outside M, sup S exists and $\sup S=\sup^M S$. Similarly, if $\inf^M S$ exists in M, then outside M, inf S exists and $\inf S=\inf^M S$.

Proof: Since "S is bounded" is absolute (for all models N such that $N\supset M$) sup S exists if and only if $\mathrm{M}\models\sup^M S$ exists. Clearly $\sup S\leq\sup^M S$. Suppose $\sup S<\sup^M S$. Then there exists a rational number t such that $\sup S<t<\sup^M S$ or $\forall r\in S\ (r<t)$. Since every rational number is in M and $\forall r\in S\ (r<t)$ and $t<\sup^M S$ are absolute, one has $\mathbf{M}\models\forall r\in S\ (r<t)\wedge t<\sup^M S$ which is a contradiction. A similar proof holds for $\inf^M S$ and $\inf S$. QED

Theorem 5: Let M be a standard transitive ZFC model. For each $P\in M$ such that $\mathrm{M}\models P$ is a probability

measure on $\mathcal{B}(\{0,1\}^{\omega}_{M})$, there is, outside **M**, a unique Q such that Q is a probability measure on $\mathcal{B}(\{0,1\}^{\omega})$ given by

$$QB_f = PB_f^M \tag{15}$$

for each code $f \in M$.

Proof: Inside **M** define a function q on the set C^M of codes by

$$qf = PB_f^M \tag{16}$$

for all $f \in C^M$. Then from the definition of codes one has (a) $f \in C_0 \rightarrow 0 \leq qf \leq 1$, (b) α odd $\rightarrow qf = 1 - qTf$, for all $f \in C_\alpha^M$, (c) α even $\rightarrow$ for all $f \in C_\alpha^M$, $qf = \sup\{qf^n \mid n = 1, 2, \cdots\}$, where f^n is the n constant code for f.

Outside M, define an auxiliary function q' on $C^M = C \cap M$ as follows: ("q is a function on the codes of M" is absolute). For each $f \in C_0$ (C_0 is absolute) let $q'f = qf$. If α is odd, let $q'f = 1 - qTf$ for each $f \in C_\alpha^M$.

Now $q = q'$ on C^M. *Proof*: For $\alpha = 0$ and α odd, $q' = q$ on C_α^M is immediate from the absoluteness of the structure of the codes. For α even one has by Lemma 3 that $\{qf^n \mid n = 1, 2, \cdots\}$ is the same set of real numbers outside **M** as inside. From Lemma 4 $\sup\{qf^n \mid n = 1, 2, \cdots\} = \sup^M\{qf^n \mid n = 1, 2, \cdots\}$ so that $q' = q$ on C_α^M. Thus outside **M**, $q = q'$ on C^M.

Outside **M** let $\mathcal{J}$ be the set of all Borel subsets of $\{0,1\}^{\omega}$ with codes in **M**. Define P' on $\mathcal{J}$ by

$$P'B_f = q'f$$

for each code f in M. By the above

$$P'B_f = PB_f^M \tag{17}$$

for each f in M. Since "e codes the union of the Borel sets coded by f and g" and "e codes the complement of the Borel set coded by f" are absolute,[5] $\mathcal{J}$ is a field of Borel sets containing the sets $B_{ni} = \{\phi \mid \phi(n) = i\}$ for all $n \in \omega$ and $i \in \{0,1\}$.

Now we claim that P' is a probability measure on $\mathcal{J}$.

Proof: First $0 \leq P'B_f \leq 1$ for each $f \in C^M$. Next, let B_f and B_g in $\mathcal{J}$ be such that $B_f \cap B_g = 0$. Then by the absoluteness of "e codes the intersection of the Borel sets coded by f and g" and "e codes the empty set"[5] one has $P'(B_f \cup B_g) = P(B_f^M \cup B_g^M) = PB_f^M + PB_g^M = P'B_f + P'B_g$. So P' is additive on $\mathcal{J}$.

Next we show P' is regular on $\mathcal{J}$. By the definition of outer regularity one has $\mathbf{M} \models P$ is outer regular $\rightarrow \mathbf{M} \models$ for all codes f $PB_f^M = \inf\{P(U) \mid U$ open and $B_f^M \subset U\} \rightarrow \mathbf{M} \models$ for all codes f $(qf = \inf\{qe \mid e$ codes an open set and $f \subset e\}$. Here Eq. (16) has been used and "$f \subset e$" denotes "the set coded by f is contained in the set coded by e". Here $\{0,1\}^{\omega}_{M}$ is considered as a topological space with the usual product topology.

Since "e codes an open set" and "$f \subset e$" are both absolute,[5,9] one has, by Lemma 4 and the definition of P' from q,

$$\mathbf{M} \models qf = \inf\{qe \mid e \text{ codes an open set and } f \subset e\} \rightarrow$$

$$qf = \inf\{qe \mid e \text{ codes an open set and } f \subset e\} \rightarrow$$

$$P'f = \inf\{qe \mid e \text{ codes an open set and } f \subset e\}.$$

By the definition of $\mathcal{J}$ one then has $\mathbf{M} \models P$ is outer regular $\rightarrow P'$ is outer regular on $\mathcal{J}$. Since, inside **M**, P is outer regular, one has the result that P' is outer regular on $\mathcal{J}$.

Application of the same argument for inner regularity ("C_e is compact" is absolute[5,9] gives the result that P' is regular on $\mathcal{J}$.

Finally application of a theorem of Alexandroff outside **M**[16] ($\{0,1\}^{\omega}$ is compact in the product topology) gives the result that P' is countably additive. Thus P' is a probability measure on $\mathcal{J}$.

Application of the Hahn extension theorem[16] gives the result that (outside **M**) there exists a unique probability measure Q on $\mathcal{B}(\{0,1\}^{\omega})$ which agrees with P' on $\mathcal{J}$. Thus Q is the desired measure and the theorem is proved. QED

Corollary 6: Let **M**, P, and Q be as in Theorem 5. Then:

(1) For each code $f \in M$, $QB_f = 0 \rightarrow \mathbf{M} \models PB_f^M = 0$.

(2) Q is a product measure $\rightarrow \mathbf{M} \models P$ is a product measure.

Proof: (1) is immediate from the theorem [Eq. (15)], and (2) follows from the theorem and the fact that the property of being a product measure is defined on cylinder sets only, all of which have codes in **M**. QED

D. Sequences and ultrafilters

Let **M** be a standard transitive ZFC model. Outside **M** let Q be a probability measure on $\mathcal{B}(\{0,1\}^{\omega})$. Q is *M-correct* for a sequence $\psi \in \{0,1\}^{\omega}$ if for each Borel set B_f with code $f \in M$, $QB_f = 1 \rightarrow \psi \in B_f$. The next theorem extends a result of Solovay[5,9] to arbitrary probability measures in M. Note that if M is countable, then there exists ψ for which Q is M-correct.

Theorem 7: Let **M**, P, and Q be as in Theorem 5 with M countable. Let $\mathcal{B}_P$ be the measure algebra constructed from P and $\{0,1\}^{\omega}_{M}$ in **M**. Then there is a canonical 1–1 correspondence between the sequences $\psi \in \{0,1\}^{\omega}$, for which Q is M correct and the set of M-generic ultrafilters on $\mathcal{B}_P$ given by

$$\psi \in B_f \rightarrow b_f \in G \tag{18}$$

for each code $f \in M$. b_f is the element of $\mathcal{B}_P$ which contains B_f^M, the Borel set in M which is coded by f.

Proof: (1) $\rightarrow$: Let Q be M correct for ψ and define G by Eq. (18). First, G is well defined. For let B_g and B_h be such that $Q(B_g \Delta B_h) = 0$ with g and $h \in M$. Then, since Q is M correct for ψ, $\psi \in B_g \rightarrow \psi \in B_h$. By Corollary 6, $P(B_g^M \Delta B_h^M) = 0$ so $B_g^M \in b_h \rightarrow B_h^M \in b_h$. Thus $b_g = b_h$.

The proof that G is an ultrafilter is immediate from the definition of an ultrafilter and is left to the reader.

To show that G is M-generic, one can proceed as follows: Let D' be a subset of $\mathcal{B}_P$ in M such that $D' \subset G$ and $D' \in M$. Since $\mathcal{B}_P$ is complete in M, $\Pi D' \in \mathcal{B}_P$. It must be shown that $\Pi D' \in G$.

Claim: There is a countable subset $D \subset D'$ with $D \in M$ such that $\Pi D = \Pi D'$. For, let $D'^{\perp}$ be the set of comple-

ments of elements of D'. Then there is a countable subset $D^{\perp}$ of $D'^{\perp}$ such that[17] $\sum D^{\perp}=\sum D'^{\perp}$. But this gives[17] $\Pi D=\Pi D'$.

Next, let $C_D^M \in M$ be a set of codes such that, for each $b \in D$, C_D^M contains exactly one code out of $\{f \mid B_f^M \in b\}$. Clearly C_D^M is countable. For each code $f \in C_D^M$ let α_f be the least ordinal such that $f \in C_{\alpha_f}^M$ [Eq. (13)].

Let β' be the smallest ordinal such that $\alpha_f < \beta'$ for each $f \in C_D^M$. If β' is even, choose $\beta=\beta'+1$; otherwise set $\beta=\beta'$. β is countable.[18] Let $g: \omega \to C_D^M$ be an enumeration of the codes in C_D^M. Since D is countable in M and $C_D^M \in M$, such a g exists in M.

By construction of the codes there exists a code $h \in C_\beta$ such that [Eq. (13)] $h(0)=1$, $(Th)(0)=2$, $(Th)_{Jm}(0)=1$, and $T(Th)_{Jm}=g(m)$ for each $m \in \omega$. By construction h codes $(\cup_m B_{g(m)}^{M\perp})^{\perp}=\cap_m B_{g(m)}^M$ and thus $b_h=\Pi_m b_{g(m)}=\Pi D$.

Since the construction of the codes is absolute, outside M, h also codes the intersection of the Borel sets coded by $g(m)$ for $m=0, 1, \cdots$. Now $b_{g(m)} \in G$ for each $m \to \psi \in B_{g(m)}$ for each $m \to \psi \in \cap_m B_{g(m)} \to \psi \in B_h \to b_h \in G \to \Pi D \in G \to \Pi D' \in G$. So G is M-generic.

(2) $\leftarrow$: Let G be an M-generic ultrafilter on β_P and define ψ by Eq. (18).

First the definition is shown to be consistent. Since G is an ultrafilter, one has for any codes f, $g \in M$:

(a) $\psi \in B_f \to b_f \in G \to b_f^{\perp} \notin G \to \psi \notin (\{0,1\}^\omega - B_f)$.

(b) $\psi \in B_f$ and $\psi \in B_g \to b_f \in G$ and $b_g \in G \to b_f \cdot b_g \in G \to \psi \in B_f \cap B_g$.

(c) $\psi \in B_f$ and $B_f \subseteq B_g \to b_f \in G$ and $b_g \in G \to \psi \in B_g$.

(d) Let D be any subset of β_P such that $D \in M$. Then $\Pi D \in \beta_P$ and, by the argument of part (1), there is a Borel set $B'^M \in \Pi D$ in M with code in M. Outside M let $E_D=\{B_f \mid f \in C_D^M\}$, where C_D^M is defined as in part (1). Then $\psi \in B_f$ for each $f \in C_D^M \to b_f \in G$ for each $f \in C_D^M \to D \subseteq G \to \Pi D \in G \to \psi \in B'=\cap E_D$, where outside M, B' has the same code that B'^M does inside M.

Finally, inside M, let $C_1^M=$ set of all codes of Borel subsets of $\{0,1\}_M^\omega$ which are sets of P measure 1. Then one has, outside M, $QB_f=1$ for each $f \in C_1^M$, where Q is the probability measure constructed from P by Theorem 5. Since, in M, $b_f=1$ for each $f \in C_1^M$ and $1 \in G$ one has, outside M, $\psi \in B_f$ for each $f \in C_1^M$. Thus Q is M-correct for ψ. QED

Corollary 8: Let Q, P, and M be as in Theorem 7. Let G and ψ be related by Eq. (18) and let Q be M correct for ψ. Then

(1) $\psi \in M[G]$ and

(2) M[G] is the smallest standard transitive ZFC model N such that $M \subseteq N$ and $\psi \in$ N.

Proof: (1) By the definition of $M[G]$ of Eq. (11) and the discussion, $G \in M[G]$ and, inside M[G], G is a M-generic ultrafilter on β_P.[5] Define ψ' inside $M[G]$ by $\psi'(n)=i \to b_{ni} \in G$. Clearly ψ' is well defined and $\psi' \in M[G]$. Since the definition of ψ' is clearly $M[G]$ absolute, one has $\psi'(n)=i \to b_{ni} \in G$ outside of $M[G]$. But Eq. (18) gives $\psi(n)=i \to b_{ni} \in G$ for each $n \in \omega$ and $i \in \{0,1\}$ so that $\psi=\psi'$ and thus $\psi \in M[G]$.

(2) Let N be such that $\psi \in N$ and $M \subseteq N$. Define, in N, G' by $b_f \in G' \to \psi \in B_f^N$ for all codes f in M. Clearly $G' \in$ N. Also, since N is transitive and "ψ is a 0—1 sequence" is absolute, $B_f^N=B_f \cap N$. Then the N absoluteness of the definition of G', Eq. (18), and $\psi \in N$ gives, outside N, $b_f \in G' \to \psi \in B_f^N \to \psi \in B_f \to b_f \in G$ for all $f \in M$ or $G'=G$. Thus $G \in$ N which gives $M[G] \subseteq N$. Finally, by part (1) $\psi \in M[G]$. QED

It is clear from this corollary that $M[\psi]=M[G]$, where $M[\psi]$ is the smallest standard transitive ZFC model N such that $M \subseteq N$ and $\psi \in N$. Also, an alternate way to prove $\psi \in M[G]$ is to exhibit the element in the Boolean model M^{β_P} which equals ψ under the map $\mathcal{J}_G$ defined by Eq. (10). In particular let ψ satisfy Eq. (8a) with $\psi(\langle \check{n}, \check{1} \rangle^{\vee})=b_{ni}$ for each $n \in \omega$ and $i \in \{0,1\}$. Then one can show that ψ satisfies Eqs. (8b) and (9) and thus that $\| \psi$ is an infinite 0—1 sequence $\|=1$ and $\|\psi \notin M\|=1$. Then by Eq. (12) $\mathbf{M}[G]\|=\mathcal{J}_G(\psi)$ is an infinite 0—1 sequence not in M. By Eqs. (10) and (18), $\langle n,i \rangle \in \mathcal{J}_G(\psi) \to \psi(\langle \check{n}, \check{i} \rangle^{\vee}) \in G \to b_{ni} \in G \to \psi \in B_{ni} \to \langle n,i \rangle \in \psi$. So $\psi=\mathcal{J}_G(\psi)$ and thus $\psi \in M[G]$.

Many of the theorems and discussions of the Sec. III are given as comparisons and correspondences "inside M" and "outside M" in the informal mathematics. Alternatively the theorems and discussion can be given as comparisons and correspondences for "inside M" and "inside N," where N is any standard transitive ZFC model such that $M \subset N$. In this case one replaces "outside M" in the theorems, proofs, and discussion by "inside N." For example Theorem 5 becomes: Let M and N be standard transitive ZFC models such that $M \subset N$. For each $P \in M$ such that inside M "P is a probability measure on $\beta(\{0,1\}_M^\omega)$" there exists, inside N, a unique Q such that Q is a probability measure on $\beta(\{0,1\}_N^\omega$ given by $QB_f^N=PB_f^M$ for each code $f \in M$. In particular one can set $N=V$ where V represents "the real universe of ZFC sets," i.e., that part of the intuitive mathematics which is axiomatizable in ZFC.

Finally a theorem is given which relates the definition of randomness, used in I and given in Sec. II, and the definition of M-correctness given above. Recall from Sec. II that a probability measure Q is correct for a sequence $\psi \in \{0,1\}^\omega$ if, for all Borel subsets B of $\{0,1\}^\omega$, B is definable from Q in $\mathcal{L}_{ZF}$ and $QB=1$, then $\psi \in B$. The following theorem generalizes a result of Solovay[19] to arbitrary probability measures on $\beta(\{0,1\}^\omega)$.

Theorem 9: Let $\mathbf{M}_0$ be the minimal standard transitive ZFC model and let Q be any nonatomic probability measure on $\beta(\{0,1\}^\omega)$ which corresponds to some measure P in M_0, by Theorem 5. Then Q is correct for $\psi \to Q$ is M_0-correct for ψ.

Proof: Let B_f be any Borel subset of $\{0,1\}^\omega$ with code f in M_0 and such that $QB_f=1$. If one can show B_f is definable in $\mathcal{L}_{ZF}$ then, by the definition of Q being correct for ψ, one has $\psi \in B_f$ and thus Q is M_0 correct for ψ.

To construct the required formula, one notes that since every element of M_0 is definable inside $\mathbf{M}_0$,[20] there is a formula θ such that $\mathbf{M}_0 \models \forall y(y=f \leftrightarrow \theta(y))$. Let θ' be the formula given by $\theta'(z) \equiv (z \in M_0 > \theta^{M_0}(z) \vee (z \notin M_0 \wedge z \neq z)$, where $\theta^{M_0}(z)$ is obtained from $\theta(z)$ by relativizing all quantifiers to M_0. Then[3] $\theta^{M_0}(z) \leftrightarrow \mathbf{M}_0 \models \theta(z)$ and one

has the result that $\forall z(z = f \rightarrow \theta'(z))$ or that θ' defines f from M_0.

But M_0 is also definable [for example, by the formula[20] $q_0(x) \equiv STM_{ZF}(x) \wedge \forall z(STM_{ZF}(z) \rightarrow x \subseteq z)$, where $STM_{ZF}(x)$ is the formula in $\mathcal{L}_{ZF}$ which says that[4] x is a standard transitive ZFC model]. Let $q(x)$ be the formula $\exists y(x \in y \wedge q_0(y)$ which expresses $x \in M_0$. Then replacement of $\forall x \in M_0$, $\exists x \in M_0$, and $z \in M_0$ in $\theta'(z)$, where x is any bound variable in $\theta'(z)$, by $\forall x q(x)$, $\exists x q(x)$, and $q(z)$, gives a formula q' which defines f by $\forall z(z = f \rightarrow q'(z))$.

Let $S(\phi)$ be the formula $\exists z q'(z) \wedge$ "z is a code and ϕ lies in the Borel set coded by z." From Eq. (13) it is clear that the expression in quotation marks is definable.[9] Thus B_f is definable by $S(\phi)$ which is thus the required formula. QED

Recall that a sequence ψ is random [M_0-random] for $Q = \otimes p$ a product measure if Q is correct [M_0-correct] for ψ.

Corollary 10: Let $Q = \otimes p$ be a product measure on $\mathcal{B}(\{0,1\}^\omega)$ with $p \in M_0$. Then ψ is random for $Q \rightarrow \psi$ is M_0 – random for Q.

Proof: Immediate from Theorem 9. QED

IV. ZFC MODELS AS CARRIERS FOR THE MATHEMATICS OF QUANTUM MECHANICS

A. Boolean valued ZFC models, states, and observables

Let M be a standard transitive ZFC model and let M be the carrier for the mathematics of quantum mechanics. Then by condition (a), Sec. II, there is, inside M, an operator algebra $B(\mathcal{H}_M)$ over a Hilbert space $\mathcal{H}_M$ such that each state preparation procedure and question observing procedure correspond, under the maps Ψ_M and Φ_M, to a density operator and projection operator in $B(\mathcal{H}_M)$.

We now work inside M. Let ρ and o be respectively a density operator and projection operator and let $p_{\rho o}$ be the probability measure on $\mathcal{B}(\{0,1\})$, the four element set of all subsets of $\{0,1\}$, defined by

$$p_{\rho o}(\{1\}) = \mathrm{Tr}_M(\rho o).$$

Let $P_{\rho o} = \otimes p_{\rho o}$ be the unique product measure on $\mathcal{B}(\{0,1\}_M^\omega)$, the set of all Borel subsets of $\{0,1\}^\omega$, and which is generated from $p_{\rho o}$. Let $\mathcal{J}_{\rho o}$ be the ideal in M of Borel subsets of $\{0,1\}_M^\omega$ of $P_{\rho o}$ measure 0, and let $\mathcal{B}_{\rho o} \equiv \mathcal{B}(\{0,1\}_M^\omega)/\mathcal{J}_{\rho o}$ be the measure algebra of equivalence classes of Borel subsets of $\{0,1\}_M^\omega$ modulo subsets of $P_{\rho o}$ measure 0. In M, $\mathcal{B}_{\rho o}$ is a complete Boolean algebra which is nonatomic if and only if $0 < p_{\rho o} < 1$.

Theorem 11: With each state ρ and question observable o in $B(\mathcal{H}_M)$ there is canonically associated a unique Boolean valued ZFC model $\mathbf{M}^{\mathcal{B}_{\rho o}}$.

Proof: $\mathbf{M}^{\mathcal{B}_{\rho o}}$ is constructed in M by the prescription of Eq. (4) from $\mathcal{B}_{\rho o}$. It is clearly unique since $\mathcal{B}_{\rho o}$ is unique. Since $\mathcal{B}_{\rho o}$ is complete one has the result that $\mathbf{M}^{\mathcal{B}_{\rho o}}$ is a Boolean values ZFC model. QED

Note that if ρ lies entirely within an eigenspace of o, then $p_{\rho o}(\{1\}) = 0$ or $= 1$ and $\mathcal{B}_{\rho o}$ reduces to 2, the two element Boolean algebra $\{0,1\}$ and the theorem becomes essentially trivial. The reason is that (see Sec. III A) $\mathbf{M}^2$ differs inessentially from the model $\mathbf{M} \equiv \{x \mid x \in M\}$ which is isomorphic to M.

Thus, if ρ is a pure state in classical mechanics,[21] theorem 11 is trivial for all questions. In quantum mechanics, Theorem 11 is trivial for only those o which commute with ρ (for ρ pure).

Note that the construction of $M^{\mathcal{B}_{\rho o}}$ depends only on the measure $p_{\rho o}$ and not on ρ and o separately. Thus, if ρ, ρ', o, and o' are such that $p_{\rho o} = p_{\rho' o'}$, then $M^{\mathcal{B}_{\rho o}} \equiv M^{\mathcal{B}_{\rho' o'}}$.

Theorem 11 can be extended in many ways which may well be relevant for quantum mechanics. We shall only briefly indicate some of these here and leave more detailed investigations to future work. The essential point is that the only requirement on $\mathcal{B}$, in the construction of $M^{\mathcal{B}}$ from M, is that it be complete in M.

For example the construction of M from $\mathcal{B}$ is not limited to infinite repetitions. It applies to any stochastic process on $\{0,1\}^\omega$, finite or infinite. Note that the theorems and results of Sec. III were given for *any* probability measure on $\mathcal{B}$ $(\{0,1\}_M^\omega)$ in M, not just product measures. In the case of n repetitions, $\mathcal{B} = \mathcal{P}(\{0,1\}^n)$ the set of all subsets of $\{0,1\}^n$, in Eq. (4). In the general infinite case $\mathcal{B}_P = \mathcal{B}(\{0,1\}_M^\omega)/\mathcal{J}_P$, where $P \in M$ is a probability measure on $\mathcal{B}(\{0,1\}_M^\omega)$.

As another example, one has for each projection operator b, the four element Boolean algebra, $\mathcal{B}_b = \{0, b, b^\perp, 1\}$ of elements of $B(\mathcal{H}_M)$, where 0 and 1 are the zero and unit operators and $b^\perp = 1 - b$. $\mathcal{B}$ is obviously complete so one can by Eq. (4) associate $\mathbf{M}^{\mathcal{B}_b}$ with b.

More generally let $\mathcal{B}$ be any complete commuting subalgebra of the projection operators in $B(\mathcal{H}_M)$. Then by Eq. (4) one can associate the Boolean valued model $\mathbf{M}^{\mathcal{B}}$ with M. Note that in both these cases as well as in the case covered in Theorem 11 one can "lift" a state ρ onto the Boolean algebra. Thus one can talk about "the probability that Q is true" for any ZF formula Q.[22]

As another example let L be a quantum logic[23] in M. That is L is a countably-complete orthomodular lattice of sets in M. For example L can be the lattice of all projection operators in $B(\mathcal{H}_M)$ in M. Since L is not Boolean, one cannot construct a Boolean valued universe directly from L by Eq. (4). However, one can proceed as follows[4,5]

The partial ordering of L induces a topology of L with basis sets $[l] = \{l' \mid l' \leq l\}$ for each l in L. A subset b of L is regular open if $b = (b^-)^0$, where b^- is the closure of b and b^0 is the interior of b. Let $\mathcal{B}_L$ be the Boolean algebra of all regular open subsets of L where $b \cdot b' = b \cap b'$, $b + b' = (b \cup b')^{-0}$, and $b^\perp = (L - b)^0$. Since $\mathcal{B}_L$ is complete in M, one can use Eq. (4) to construct a Boolean valued ZFC model $M^{\mathcal{B}_L}$ from L in M.

In this manner one can assign one Boolean valued ZFC model to the whole logic rather than assigning a separate one to each complete Boolean subalgebra of L as in the previous examples. This construction, applied to partially ordered sets, is also the method by

which one shows the equivalence of the construction of Cohen extensions of M by the method of Boolean valued models and the construction by the method of forcing which uses, directly, partially ordered sets.[2,4,5]

B. Cohen extensions of M_0 and infinite repetitions

Let $\mathbf{M}_0$ be the minimal standard transitive ZFC model. Recall from I that from conditions (a) and (b) (given in Sec. II) and a strong definition of randomness (the one given here in Sec. II), $\mathbf{M}_0$ was shown not to be suitable as a carrier for the mathematics of quantum mechanics, in that no infinite outcome sequence belonged to M_0. Here condition (a) only will be considered as holding in $\mathbf{M}_0$. Then condition (b) will be used outside M_0 [i.e., by dropping all references to $\mathbf{M}$ in condition (b)] to extend M_0.

The results of the previous sections can be combined into the following theorem:

Theorem 12: Let Ψ_{M_0} and Φ_{M_0} be as in condition (a) and assume that the definition of randomness given in Sec. II is correct. Then to each state preparation procedure $s \in \mathrm{Dom}(\Psi_{M_0})$ and question measuring procedure $q \in \mathrm{Dom}(\Phi_{M_0})$ such that $\Psi_{M_0}(s)$ is not dispersion free for $\Phi_{M_0}(q)$, and to each $t:\omega \to R$ in M_0 such that (tsq) corresponds to an infinite repetition of carrying out s followed by q, there corresponds a unique standard transitive ZFC model $\mathbf{M}_0[\psi_{tsq}]$, where ψ_{tsq} is the random outcome sequence associated with (tsq) by condition (b). $\mathbf{M}_0[\psi_{tsq}]$ is the smallest standard transitive ZFC model $\mathbf{N}$ such that $M_0 \subseteq N$ and $\psi_{tsq} \in N$.

Proof: Inside $\mathbf{M}_0$ let p_{sq} be the probability measure on the set of all subsets of $\{0,1\}$ given by $p_{sq}(\{1\}) = \mathrm{Tr}_{M_0}(\Psi_{M_0}(s)\Phi_{M_0}(q))$ and in $\mathbf{M}_0$ let $P_{sq} = \otimes p_{sq}$ be the product probability measure constructed from p_{sq}. Since $\Psi_{M_0}(s)$ is not dispersion free for $\Phi_{M_0}(q)$, P_{sq} is nonatomic. By Theorem 11, one has the unique Boolean valued ZFC model $\mathbf{M}_0^{\beta_{sq}}$, where β_{sq} is the nonatomic measure algebra constructed from $\beta(\{0,1\}^{\omega}_{M_0})$ and P_{sq}.[15]

Outside $\mathbf{M}_0$ let Q_{sq} be the probability measure on $\beta(\{0,1\}^{\omega})$ which corresponds to P_{sq} according to Theorem 5 and Eq. (15). By Corollary 6 Q_{sq} is a nonatomic product measure and in fact $Q_{sq} = \otimes p_{sq}$ since "p_{sq} is a probability measure on $\beta(\{0,1\})$" is absolute. By condition (b) applied outside $\mathbf{M}_0$ and the definition of randomness given in Sec. II, there exists an outcome sequence ψ_{tsq} such that Q_{sq} is correct for ψ_{tsq}. By Theorems 9 and 7 there is a unique M_0-generic ultrafilter G_{tsq} on β_{sq} given by Eq. (18). By Eqs. (10) and (11) one defines $M_0[G_{tsq}]$ which, by Corollary 8, $= M_0[\psi_{tsq}]$ and is the smallest standard transitive ZFC model $\mathbf{N}$ such that $M_0 \subseteq N$ and $\psi_{tsq} \in N$.

There are several aspects of these results worth noting. First Theorem 11 holds for any standard transitive ZFC model, whereas Theorem 12 holds only for those models $\mathbf{M}$ for which Theorem 9 (with M replacing M_0) holds. Furthermore, Theorem 12 clearly depends on which definition of randomness one uses. For definitions for which Theorem 9 holds, such as that of Solovay[9] (generalized to arbitrary product measures on $\{0,1\}^{\omega}$), Theorem 12 holds. For weaker definitions such as those of Martin Löf,[13] $\psi_{tsb} \in M_0$ and Theorem 12 fails.

If s and q are such that $\Psi_{M_0}(s)$ is dispersion free for $\Phi_{M_0}(q)$, then either $p_{sq}(\{1\}) = 1$ or $p_{sq}(\{1\}) = 0$ and for each infinite repetition (tsq), ψ_{tsq} is the constant 1 sequence or the constant 0 sequence. In this case β_{sq} is two element Boolean algebra $\{0,1\}$ and the only possible M_0-generic ultrafilter on β_{sq} is $\{1\}$, which lies in M_0. Carrying through the construction of Eqs. (10) and (11) gives one $\mathbf{M}_0$ again with $\psi_{tsq} \in M_0$, so that Theorem 12 becomes trivial.

The difference between classical and quantum mechanics, noted before in I, appears here again. For classical mechanics based on $\mathbf{M}_0$, CM_{M_0}, the range of Φ_{M_0} is a Boolean algebra. Thus for each *pure* state $\Psi_{M_0}(s)$ in CM_{M_0}, Theorem 12 is trivial and the construction does not lead outside M_0 for *all question procedures* q. This is not the case of QM_{M_0}, where, for each s such that $\Psi_{M_0}(s)$ is pure, there exist question procedures q such that $[\Psi_{M_0}(s), \Phi_{M_0}(q)] \neq 0$ [i.e., $\Psi_{M_0}(s)$ is not dispersion free for $\Phi_{M_0}(q)$].

Theorem 12 as given refers to infinite repetitions of measurements which generate probability measures on $\beta(\{0,1\}^{\omega})$ in $\mathbf{M}_0$, which are product measures. However, the results of Sec. III are not so limited. Thus these results can be used to give extensions of Theorems 11 and 12 which apply to general stochastic processes in quantum mechanics and not just infinite repetitions. Examples of these are sequences of successive measurements on the same system. Of course, conditions (a) and (b) must be extended to include such processes.

V. ELIMINATION OF ZFC MODELS M_0 $[\psi_{tsq}]$

The arguments used in I to exclude $\mathbf{M}_0$ as a possible mathematical universe for quantum mechanics cannot be used directly to exclude the models $\mathbf{M}_0[\psi_{tsq}]$. The reason is that condition (b), which requires that ψ_{tsq} be random and $\psi_{tsq} \in N$, clearly holds if $N = M_0[\psi_{tsq}]$. [The failure of condition (b) for $\mathbf{N} = \mathbf{M}_0$ is the essential part of the proof of Theorem 2.]

However, the arguments can be extended to include the models $\mathbf{M}_0[\psi_{tsq}]$, and it is the aim of this section to show that no ZFC model $\mathbf{M}_0[\psi_{tsq}]$ can serve as the mathematical universe for quantum mechanics.

One must consider here another necessary condition a ZFC model $\mathbf{M}$ must satisfy if it is to serve as the mathematical universe for quantum mechanics. Let s and u be two different state preparation procedures, and let q and k be two different question measuring procedures, and let ψ_{tsq} and ψ_{wuk} be the respective outcome sequences associated with the infinite repetitions (tsq) and (wuk) of doing s and q, and u and k.

Intuitively one requires that it be impossible to predict any outcome of carrying out u and k, given prior knowledge of the state $\Psi_M(u)$ and observable $\Phi_M(k)$. This is taken care of by requiring that ψ_{wuk} be random (i.e., that P_{uk} be correct for ψ_{wuk}). However, one also requires that it be impossible to predict any outcome of carrying out u and k given the additional prior knowledge of ψ_{tsq} for any infinite repetition (tsq). Similarly one requires that it be impossible to predict any outcome of doing s and q given prior knowledge of ψ_{wuk}. In a word

one requires that ψ_{tsq} and ψ_{wuk} be mutually statistically independent.

On intuitive grounds, then, this requirement of independence should be included as another condition which a ZFC model **M** must satisfy if it is to serve as the mathematical universe for quantum mechanics. That is one has the following necessary condition (c):

(c): For each pair s and u of different state preparation procedures in the domain of Ψ_M and for each pair q and k of different question measuring procedures in the domain of Φ_M and for each infinite repetition (wuk) of doing u and k and for each infinite repetition (tsq) of doing s and q, the outcome sequences ψ_{wuk} and ψ_{tsq}, which are associated with (wuk) and (tsq) by condition (b), are mutually statistically independent.

In keeping with the definition of randomness used here, the following definition of independence is reasonable. A sequence $\psi \in \{0, 1\}^\omega$ is *independent* of a sequence ψ' if ψ is not definable from ψ'. That is, for no formula Q in $\mathcal{L}_{ZF}$, the language of ZF set theory does $\forall x(x = \psi \leftrightarrow Q(x, \psi'))$ hold. ψ and ψ' are *mutually statistically independent* if ψ is independent of ψ' and ψ' is independent of ψ.

A product probability measure $Q = \otimes p$ on $\beta(\{0, 1\}^\omega)$ outside $\mathbf{M}_0$ is an M_0 *product measure* if $p \in M_0$. Also one says that ψ is *random for* Q if Q is correct for ψ (Sec. II).

Lemma 13: Let $Q = \otimes p$ be a nonatomic M_0 product measure and let ψ be random for Q. Let $\mathbf{M}_0[\psi]$ be the smallest standard transitive ZFC model **N** such that $M_0 \subseteq N$ and $\psi \in N$ (Sec. III, Corollary 8). If ψ' is statistically independent of ψ, then $\psi' \notin M_0[\psi]$.

Proof: Assume the converse, i.e., that ψ' is statistically independent of ψ and $\psi' \in M_0[\psi]$.

(1): There exists a formula $\theta(x, \psi, M_0, c)$ which defines ψ' from ψ, M_0, and some $c \in M_0$, i.e., $\forall x(x = \psi' \leftrightarrow \theta(x, \psi, M_0, c)$. To see this, one first notes that, by Corollary 8 and Eqs. (10) and (11), there is some $d \in M_0^{\beta_P}$ (in M_0, $P = \otimes p$) such that $\mathcal{J}_{G_\psi}(d) = \psi'$. In particular $\mathcal{J}_{G_\psi \gamma+1}(d) = \psi'$, where $\gamma = \text{rank } d =$ least ordinal such that $d \in M_{0\gamma+1}^{\beta_P}$ [Eq. (4)] and G_ψ is defined on β_P by Eq. (18). It is clear from Eqs. (4) and (10) that $\mathcal{J}_{G_\psi \gamma+1}$ is definable from P, M_0, ψ, and d (as a function on $M_{0\text{rank}(d)+1}^{\beta_P}$). Let $Q_1(x, M_0, \psi, P, d)$ be the formula which defines $\mathcal{J}_{G_\psi \text{rank} d}$ [by $\forall x(x = \mathcal{J}_{G_\psi \text{rank}(d)+1} \leftrightarrow Q_1(x, M_0, \psi, P, d)$].

Let c be the ordered pair $\langle d, P\rangle$. Then[3,5] $c = \langle d, P\rangle$ inside M_0. Then $\exists y(\langle l(c), x\rangle \in y \wedge Q_1(y, M_0, \psi, r(c), l(c)))$ is the desired formula $\theta(x, \psi, M_0, c)$, where $l(c)$ and $r(c)$ denote the respective left hand and right-hand elements of c.

(2): By the same argument as was used in the proof of theorem 9, there is a formula $Q'(y)$ in $\mathcal{L}_{ZF}$ which defines c [i.e., $\forall y(y = c \leftrightarrow Q'(y))$]. Thus the formula $Q'_{M_0}(x, \psi) \equiv \exists y(\theta(x, \psi, M_0, y) \wedge Q'(y))$ defines ψ' from ψ and M_0. Let $\theta_0(x, \psi)$ be the formula obtained from $Q'_{M_0}(x, \psi)$ by replacing each occurence of $w = M_0$ by $q_0(w)$ (see proof of Theorem 9) and each occurrence of $w \in M_0$ by $\exists z(w \in z \wedge q_0(z))$, where w is any variable, other than x, in $Q'_{M_0}(x, \psi)$. Then $\theta_0(x, \psi)$ defines ψ' from ψ.

But this contradicts the statistical independence of ψ' from ψ. Thus $\psi' \notin M_0[\psi]$. QED

Corollary 14: Let ψ and ψ' both be random for nonatomic M_0 product measures and let ψ and ψ' be mutually statistically independent. Then $\psi \notin M_0[\psi']$ and $\psi' \notin M_0[\psi]$.

Proof: Immediate from Lemma 13 and the definition of mutual statistical independence. QED

In Theorem 12 it was seen that with each $s \in \text{Dom}(\Psi_{M_0})$ and $q \in \text{Dom}(\Phi_{M_0})$ for which $\Psi_{M_0}(s)$ is not dispersion free for $\Phi_{M_0}(q)$, (i.e., $0 < p_{sq}(\{1\}) = \text{Tr}_{M_0}(\Psi_{M_0}(s)\Phi_{M_0}(q)) < 1$) and with each $t \in M_0$ with $t : \omega \to R_{M_0}$ (inside $\mathbf{M}_0$) such that (tsq) is an infinite repetition of doing s and q at times $t(0), t(1), \cdots$ there is associated a unique standard transitive ZFC model $\mathbf{M}_0[\psi_{tsq}]$, where ψ_{tsq} is the random outcome sequence associated with (tsq).

Now consider $\mathbf{M}_0[\psi_{tsq}]$ as a possible carrier for the mathematics of quantum mechanics. If $\mathbf{M}_0[\psi_{tsq}]$ is satisfactory, then conditions (a), (b), and (c) must hold for $QM_{M_0[\psi_{tsq}]}$, i.e., quantum mechanics based in $\mathbf{M}_0[\psi_{tsq}]$. By Theorem 1 and Eqs. (2a) and (3a),

$$\begin{aligned} &\text{Dom}(\Phi_{M_0}) \subseteq \text{Dom}(\Phi_{M_0[\psi_{tsq}]}), \\ &\text{Dom}(\Psi_{M_0}) \subseteq \text{Dom}(\Psi_{M_0[\psi_{tsq}]}), \\ &\text{Tr}_{M_0[\psi_{tsq}]}(\Psi_{M_0[\psi_{tsq}]}(s')\Phi_{M_0[\psi_{tsq}]}(q')) \\ &\quad = \text{Tr}_{M_0}(\Psi_{M_0}(s')\Phi_{M_0}(q')) \end{aligned} \tag{19}$$

holds for each $s' \in \text{Dom}(\Psi_{M_0})$ and $q' \in \text{Dom}(\Phi_{M_0})$.

Let $u \neq s$ be a state preparation procedure in $\text{Dom}(\Psi_{M_0})$ and $k \neq q$ be a question measuring procedure in $\text{Dom}(\Phi_{M_0})$ such that $\Psi_{M_0}(u)$ is not dispersion free for $\Phi_{M_0}(k)$. Then by the above $u \in \text{Dom}(\Psi_{M_0[\psi_{tsq}]})$ and $k \in \text{Dom}(\Phi_{M_0[\psi_{tsq}]})$ and the measure $P_{uk} = \otimes p_{uk}$ in $M_0[\psi_{tsq}]$ is a nonatomic product measure on $\beta(\{0, 1\}^\omega_{M_0[\psi_{tsq}]})$. Also for each $w \in M_0$ such that, inside M_0, $w : \omega \to R_{M_0}$ is increasing, one has, by absoluteness and $M_0 \subseteq M_0[\psi_{tsq}]$, that $w \in M_0[\psi_{tsq}]$ and, inside $M_0[\psi_{tsq}]$, $w : \omega \to R_{M_0[\psi_{tsq}]}$ is increasing.

For each $w \in M_0$ with $w : \omega \to R_{M_0}$ such that (wuk) is an infinite repetition of doing u and k, one has from conditions (a) and (b) for $QM_{M_0[\psi_{tsq}]}$ that ψ_{wuk} is random for P_{uk} and $\psi_{wuk} \in M_0[\psi_{tsq}]$. By the construction, and Eq. (19), in $QM_{M_0[\psi_{tsq}]}$, (tsq) is an infinite repetition of carrying out s and q, P_{sq} is a nonatomic product measure, and ψ_{tsq} is random for P_{sq} with $\psi_{tsq} \in M_0[\psi_{tsq}]$.

By condition (c) ψ_{wuk} is statistically independent of ψ_{tsq}. Thus, if the definition of statistical independence given here is correct, then by Lemma 13 $\psi_{wuk} \notin M_0[\psi_{tsq}]$, which is a contradiction. Thus condition (b) is violated and $\mathbf{M}_0[\psi_{tsq}]$ cannot serve as the carrier for the mathematics of quantum mechanics. Exchanging (tsq) and (wuk) gives the result that $\mathbf{M}_0[\psi_{wuk}]$ also cannot be a carrier.

Thus the following theorem has been proved.

Theorem 15: Let the definitions of randomness and statistical independence given here be correct. Then for each $s \in \text{Dom}(\Psi_{M_0})$ and $q \in \text{Dom}(\Phi_{M_0})$ and $t : \omega \to R_{M_0}$ with $t \in M_0$ such that (tsq) is an infinite repetition of

doing s and q, $\mathbf{M}_0[\psi_{tsq}]$ cannot be a carrier for the mathematics of quantum mechanics.

VI. DISCUSSION

The exclusion of ZFC models of the type $\mathbf{M}_0[\psi_{tsq}]$ as possible carriers for the mathematics of quantum mechanics, clearly depends on the strength of the definition of statistical independence which is used. This fact, which is expressed explicitly in Theorem 15, is similar to the situation which obtains for definitions of randomness as discussed in I.

In fact, for each definition of randomness, one has a corresponding definition of statistical independence. For example the definition of randomness used by Solovay,[9] generalized to arbitrary product measures and applied to M_0 is as follows: A sequence ψ is S-M_0 random if there exists an M_0 product measure Q on $\mathcal{B}(\{0,1\}^\omega)$ such that for all Borel sets B if B has a code [Eq. (13)] in M_0 and $QB=1$, then $\psi\in B$. The corresponding definition of statistical independence is as follows: A sequence ψ is S-M_0 statistically independent of a sequence ψ' if the Borel set $\{\psi\}$ has no code in $M_0[\psi']$. (Note that if ψ is S-$M_0[\psi']$ random for a nonatomic measure, then ψ is S-M_0 random and ψ is S-M_0 statistically independent of ψ'.)

These definitions of statistical independence and randomness are weaker[24] than the definitions used here. However, they are sufficiently strong so that Theorem 2 and 15 hold for them also. That is, for these definitions, neither $\mathbf{M}_0$ nor $\mathbf{M}_0[\psi_{tsq}]$, where (tsq) is any infinite repetition in QM_{M_0} and P_{sq} is nonatomic, are suitable as carriers for the mathematics of quantum mechanics.

One can also give definitions of statistical independence which correspond to the two definitions of randomness given by Martin Löf.[13] These definitions are too weak for Theorems 2 and 15 to hold as, inside M_0, there exist Martin Löf random sequences. Also the corresponding definition of statistical independence can be applied inside $\mathbf{M}_0$.

As noted in I an important open question is to investigate whether or not the definition of randomness must be at least as strong as (—). Here one sees that this question also includes the definition of statistical independence. Thus one would like to be able to prove that the definitions of randomness and statistical independence are at least as strong as (—).

It is speculated that such a proof will not be forthcoming until one axiomatizes physics and mathematics together in one coherent theory instead of treating them separately, as has been done so far. Such treatment will probably include the observer more intimately than has been done so far. In this connection note that one can regard the different definitions of randomness and independence as corresponding to different predictive powers of an observer.

Finally one notes that in the usual ZFC model V there exist sequences which are random and statistically independent for any of the above definitions. Thus one might argue that one should just take QM_V, i.e., quantum mechanics based on the real ZFC world V (as has been implicitly done in all of physics so far) and ignore quantum mechanics based on other ZFC models M. The point to be made is that why the real ZFC world (assuming that such exists) is V and not some other model M needs explaining. An implicit point of this and the preceeding paper is that such an explanation may be forthcoming only from a coherent theory of physics and mathematics. More explicitly, it has been shown here and in I that randomness and statistical independence may have a direct bearing on this problem.

ACKNOWLEDGMENT

The author is indebted to Robert Solovay for helpful comments, especially on the relationship between the definitions of randomness of Ref. 9 and that used here.

*Based on work performed under the auspices of the U.S. Energy Research and Development Administration.

[1]Paul Benioff, J. Math. Phys. 17, 618 (1976).
[2]Paul J. Cohen, *Set Theory and the Continuum Hypothesis* (Benjamin, New York, 1966).
[3]Gaisi Takeuti and Wilson Zaring, *Introduction to Axiomatic Set Theory* (Springer, New York, 1971).
[4]Gaisi Takeuti and Wilson Zaring, *Axiomatic Set Theory* (Springer, New York, 1973).
[5]Thomas Jech, *Lectures in Set Theory*, Mathematics Lecture Notes #217 (Springer-Verlag, Berlin, 1971).
[6]Ulrich Felgner, *Models of ZF Set Theory*, Mathematics Lecture Notes #223 (Springer-Verlag, Berlin, 1971).
[7]J. Barkley Rosser, *Simplified Independence Proofs* (Academic, New York, 1969).
[8]R.M. Solovay and S. Tennenbaum, Ann. Math. 94, 201 (1971).
[9]R.M. Solovay, Ann. Math. 92, 1 (1970).
[10]Leon Henkin, Amer. Math. Monthly 78, 463 (1971) and A.A. Fraenkel, Y. Bar-Hillel, and A. Levy, *Foundations of Set Theory*, 2nd rev. ed. (North-Holland, Amsterdam, 1973), Chap. 5, Sec. 9.
[11]P.A. Benioff, Phys. Rev. D 7, 3603 (1973).
[12]A.H. Kruse, Z. Mat. Logik Grundlagen Math. 13, 299 (1967).
[13]Per Martin Löf, *On the Notion of Randomness*, Proceedings of Summer Institute on Proof Theory and Intuitionism, State University of New York, Buffalo, 1968, edited by J. Myhill, A. Kino and R. Vesley (North-Holland, Amsterdam, 1970); Inform. Control 9, 603 (1966).
[14]M^β is our first example of a nonstandard ZFC model.
[15]Paul Halmos, *Lectures on Boolean Algebra* (Van Nostrand, Princeton, N.J., 1963), Chap. 15.
[16]Nelson Dunford and Jacob Schwartz, *Linear Operators* (Wiley Interscience, New York), Vol. 1, 136—138.
[17]Reference 15, pp. 26, 61.
[18]Felix Hausdorff, *Set Theory* (Chelsea Publishing Co., New York, 1962), 2nd ed. p. 84.
[19]Robert Solovay, private communication.
[20]Y. Suzuki and G. Wilmers, *Non-Standard Models for Set Theory*, Proceedings of the Bertrand Russell Memorial Logic Conference, Denmark, 1971, edited by John Bell, Julian

Cole, Graham Priest and Alan Slomsen (Bertrand Russell Memorial Logic Conference, Leeds, 1973), pp. 278—314.

[21]A corresponding construction for classical mechanics is as follows: Let $S_M \in \mathbf{M}$ be the phase space of a physical system and $A_M \in M$ be such the $\mathbf{M} \models A_M$ is the algebra of all continuous real valued functions on S_M. Each state corresponds to a $\rho \in M$ such that $\mathbf{M} \models \rho$ is a regular Borel measure on $\mathcal{B}(S_M)$, the set of all Borel subsets of S_M. Each question observable corresponds to a characteristic function χ_o with $p_{\rho o}(\{1\}) = \int_{S_M} \chi_o \, dp$, inside $\mathbf{M}$. One then constructs $\mathcal{B}_{\rho o}$ and $\mathbf{M}\mathcal{B}_{\rho o}$ as in the text.

[22]Dana Scott and Peter Krauss, Assigning Probabilities to Logical Formulas, in *Aspects of Inductive Logic*, edited by J. Hintikka and P. Suppes (North-Holland, Amsterdam, 1966), pp. 219—64.

[23]V. S. Varadarajan, *Geometry of Quantum Theory* (Van Nostrand, Princeton, N.J., 1968), Vol. Chap. VI.

[24]For randomness, see Corollary 10 and Sec. IV of Ref. 1. For statistical independence one replaces M_0 by $M_0[\psi]$ in Theorem 9 and Corollary 10 and carries out the proof as is done in Lemma 13.

Conformal transformation for the Coulter–Weinberg form of the equations for mass zero spin-2 field

Carlos D. Galles* and Oscar S. Zandron

Departamento de Física, Facultad de Ciencias Exactas e Ingeniería, Universidad Nacional de Rosario, Rosario, Argentina
(Received 15 October 1975)

In this work we show that it is impossible to introduce a third-rank tensor potential that preserves the conformal covariance of the mass zero spin-2 field equations in the Coulter–Weinberg scheme.

1. INTRODUCTION

The description of a field of zero mass and spin 2 and of their electromagnetic and gravitational interactions has been the object of a large number of researches. These researches take as a starting point a Lagrangian formulation, as in Fierz and Pauli's original work,[1] or they use the irreducible unitary representations of Poincarê's group,[2] as in Weinberg's works.[3]

From Weinberg's research, Coulter[4] develops a formulation for the field of zero mass and spin 2, the equations of which are expressed according to the ten "physical components" of the free field.

By writing these equations in a form manifestly covariant under the Lorentz group, it is easy to see that they are also formally covariant under the group of conformal transformations. It is convenient to emphasize that this group is the one of largest dimension that preserves the nullity of the element of line[5]; this gives place to a symmetry in the Minkowsky space required to the particles of zero mass.

The introduction of potentials according to which the physical components of the field can be written, a necessary step for the study of the interacting field, presents some difficulties.

In this work a negative answer is given to the existence of suitable potentials of the field that preserve the conformal covariance according to Coulter's scheme.

2. NONEXISTENCE OF A SATISFACTORY THIRD RANK TENSOR POTENTIAL

In the Lorentz manifestly covariant formulation of the field of zero mass and spin 2 developed by Coulter,[4] a fourth rank tensor $R_{\mu\nu\lambda\sigma}$ that verifies the Weyl's tensor conditions is used:

$$R_{\mu\nu\lambda\sigma} = -R_{\mu\nu\sigma\lambda}, \quad R_{\mu\nu\lambda\sigma} = R_{\lambda\sigma\mu\nu}, \tag{1}$$

$$R_{\mu\lambda} \equiv \eta^{\nu\sigma} R_{\mu\nu\lambda\sigma} = 0, \tag{2}$$

$$\epsilon^{\mu\nu\lambda\sigma} R_{\mu\nu\lambda\sigma} = 0. \tag{3}$$

The dynamics of the field in interaction is given by the equation

$$\partial^{\mu} R_{\mu\nu\lambda\sigma} = \kappa \mathcal{J}_{\nu\lambda\sigma}, \tag{4}$$

where $\mathcal{J}_{\nu\lambda\sigma}$ is the source of the field and κ is a coupling constant.

In his paper Coulter found the conditions for the covariance of these equations under C^+, the connected component of the conformal group C. These conditions are

$$R'_{\mu\nu\lambda\sigma}(x') = \mathcal{J}^{-1} \frac{\partial x^{\alpha}}{\partial x'^{\mu}} \frac{\partial x^{\beta}}{\partial x'^{\nu}} \frac{\partial x^{\gamma}}{\partial x'^{\lambda}} \frac{\partial x^{\delta}}{\partial x'^{\sigma}} R_{\alpha\beta\gamma\delta}(x), \tag{5}$$

$$\mathcal{J}'_{\nu\lambda\sigma}(x') = \mathcal{J} \frac{\partial x^{\beta}}{\partial x'^{\nu}} \frac{\partial x^{\gamma}}{\partial x'^{\lambda}} \frac{\partial x^{\delta}}{\partial x'^{\sigma}} \mathcal{J}_{\beta\gamma\delta}(x). \tag{6}$$

He also showed that the results obtained fail all in the preservation of the conformal covariance when it is written $R_{\mu\nu\lambda\sigma}$ in terms of: (1) a second rank tensor potential; (2) a symmetrical second rank tensor potential; (3) a third rank tensor potential symmetrical in the last two indexes, containing the sum of the irreducible representations $(\frac{3}{2},\frac{3}{2})$, $(\frac{3}{2},\frac{1}{2})$, $(\frac{1}{2},\frac{1}{2})$, $(\frac{1}{2},\frac{1}{2})$ and $(\frac{1}{2},\frac{3}{2})$. Nevertheless, there could be a number of other possibilities.

We prove that the result is equally negative, if we use another potential of third rank $H_{\mu\nu\lambda}$, antisymmetrical in the last two indexes and satisfying the condition $\epsilon^{\mu\nu\lambda\rho} H_{\mu\nu\lambda} = 0$. This tensor has 20 independent components, which transform according to the representation $(\frac{3}{2},\frac{1}{2}) \oplus (\frac{1}{2},\frac{1}{2}) \oplus (\frac{1}{2},\frac{3}{2})$, this being the minimum number of irreducible representations that can describe the spin 2.[6]

This sequence of negative results leads us to formulate the problem in a general way. For that it is necessary to write $R_{\mu\nu\lambda\sigma}$ as a function of a third rank tensor $T_{\mu\nu\lambda}$ without any condition "*a priori*"; the development takes the following form:

$$\begin{aligned} R_{\mu\nu\lambda\sigma} = {} & a(\partial_{\mu} T_{\nu\lambda\sigma} + \partial_{\lambda} T_{\sigma\mu\nu} - \partial_{\mu} T_{\nu\sigma\lambda} - \partial_{\lambda} T_{\sigma\nu\mu} + \partial_{\sigma} T_{\lambda\nu\mu} + \partial_{\nu} T_{\mu\sigma\lambda} - \partial_{\sigma} T_{\lambda\mu\nu} - \partial_{\nu} T_{\mu\lambda\sigma}) \\ & + b(\partial_{\mu} T_{\lambda\nu\sigma} + \partial_{\lambda} T_{\mu\sigma\nu} - \partial_{\mu} T_{\sigma\nu\lambda} - \partial_{\lambda} T_{\nu\sigma\mu} + \partial_{\sigma} T_{\nu\lambda\mu} + \partial_{\nu} T_{\sigma\mu\lambda} - \partial_{\sigma} T_{\mu\lambda\nu} - \partial_{\nu} T_{\lambda\mu\sigma}) \\ & + c(\partial_{\mu} T_{\lambda\sigma\nu} + \partial_{\lambda} T_{\mu\nu\sigma} - \partial_{\mu} T_{\sigma\lambda\nu} - \partial_{\lambda} T_{\nu\mu\sigma} + \partial_{\sigma} T_{\nu\mu\lambda} + \partial_{\nu} T_{\sigma\lambda\mu} - \partial_{\sigma} T_{\mu\nu\lambda} - \partial_{\nu} T_{\lambda\sigma\mu}) \\ & - \tfrac{1}{2}(b+c)[\eta_{\mu\lambda}(\partial_{\nu} T_{\sigma}{}^{\rho}{}_{\rho} + \partial_{\sigma} T_{\nu}{}^{\rho}{}_{\rho}) - \eta_{\nu\lambda}(\partial_{\mu} T_{\sigma}{}^{\rho}{}_{\rho} + \partial_{\sigma} T_{\mu}{}^{\rho}{}_{\rho}) + \eta_{\nu\sigma}(\partial_{\mu} T_{\lambda}{}^{\rho}{}_{\rho} + \partial_{\lambda} T_{\mu}{}^{\rho}{}_{\rho}) - \eta_{\mu\sigma}(\partial_{\nu} T_{\lambda}{}^{\rho}{}_{\rho} + \partial_{\lambda} T_{\nu}{}^{\rho}{}_{\rho})] \end{aligned}$$

$$+\tfrac{1}{2}(a+b)[\eta_{\mu\lambda}(\partial_\nu T^\rho{}_{\rho\sigma}+\partial_\sigma T^\rho{}_{\rho\nu})-\eta_{\nu\lambda}(\partial_\mu T^\rho{}_{\rho\sigma}+\partial_\sigma T^\rho{}_{\rho\mu})+\eta_{\nu\sigma}(\partial_\mu T^\rho{}_{\rho\lambda}+\partial_\lambda T^\rho{}_{\rho\mu})-\eta_{\mu\sigma}(\partial_\nu T^\rho{}_{\rho\lambda}+\partial_\lambda T^\rho{}_{\rho\nu})]$$

$$+\tfrac{1}{2}(c-a)[\eta_{\mu\lambda}(\partial_\nu T^\rho{}_{\sigma\rho}+\partial_\sigma T^\rho{}_{\nu\rho})-\eta_{\nu\lambda}(\partial_\mu T^\rho{}_{\sigma\rho}+\partial_\sigma T^\rho{}_{\mu\rho})+\eta_{\nu\sigma}(\partial_\mu T^\rho{}_{\lambda\rho}+\partial_\lambda T^\rho{}_{\mu\rho})-\eta_{\mu\sigma}(\partial_\nu T^\rho{}_{\lambda\rho}+\partial_\lambda T^\rho{}_{\nu\rho})]$$

$$+\tfrac{1}{2}(c-a)[\eta_{\mu\lambda}(\partial^\rho T_{\nu\rho\sigma}+\partial^\rho T_{\sigma\rho\nu})-\eta_{\nu\lambda}(\partial^\rho T_{\mu\rho\sigma}+\partial^\rho T_{\sigma\rho\mu})+\eta_{\nu\sigma}(\partial^\rho T_{\mu\rho\lambda}+\partial^\rho T_{\lambda\rho\mu})-\eta_{\mu\sigma}(\partial^\rho T_{\nu\rho\lambda}+\partial^\rho T_{\lambda\rho\nu})]$$

$$-\tfrac{1}{2}(b+c)[\eta_{\mu\lambda}(\partial^\rho T_{\rho\nu\sigma}+\partial^\rho T_{\rho\sigma\nu})-\eta_{\nu\lambda}(\partial^\rho T_{\rho\mu\sigma}+\partial^\rho T_{\rho\sigma\mu})+\eta_{\nu\sigma}(\partial^\rho T_{\rho\mu\lambda}+\partial^\rho T_{\rho\lambda\mu})-\eta_{\mu\sigma}(\partial^\rho T_{\rho\nu\lambda}+\partial^\rho T_{\rho\lambda\nu})]$$

$$+\tfrac{1}{2}(a+b)[\eta_{\mu\lambda}(\partial^\rho T_{\nu\sigma\rho}+\partial^\rho T_{\sigma\nu\rho})-\eta_{\nu\lambda}(\partial^\rho T_{\mu\sigma\rho}+\partial^\rho T_{\sigma\mu\rho})+\eta_{\nu\sigma}(\partial^\rho T_{\mu\lambda\rho}+\partial^\rho T_{\lambda\mu\rho})-\eta_{\mu\sigma}(\partial^\rho T_{\nu\lambda\rho}+\partial^\rho T_{\lambda\nu\rho})]$$

$$+\tfrac{2}{3}(\eta_{\mu\lambda}\eta_{\nu\sigma}-\eta_{\nu\lambda}\eta_{\mu\sigma})[(b+c)\partial^\rho T_\rho{}^\tau{}_\tau+(a-c)\partial^\rho T^\tau{}_{\rho\tau}-(a+b)\partial^\rho T^\tau{}_{\tau\rho}]. \tag{7}$$

The expression (7) verifies (1) and (2), being a, b, and c undeterminate constants. The restrictions (3) supply the condition

$$T_{\nu\lambda\sigma}+T_{\sigma\nu\lambda}+T_{\lambda\sigma\nu}-T_{\nu\sigma\lambda}-T_{\sigma\lambda\nu}-T_{\lambda\nu\sigma}=0. \tag{8}$$

To transform $R_{\mu\nu\lambda\sigma}$ covariantly under C^* and according to (5), it is necessary to satisfy the following equations:

$$(a+b-c)T_{\nu\lambda\sigma}+(-a-b-c)T_{\sigma\nu\lambda}+(-a+b+c)T_{\lambda\sigma\nu}+(-a-b+c)T_{\nu\sigma\lambda}+(a-b-c)T_{\sigma\lambda\nu}+(a+b+c)T_{\lambda\nu\sigma}=0, \tag{9a}$$

$$(-a+c)T_{\nu\lambda\sigma}+(a+b)T_{\sigma\nu\lambda}+(-b-c)T_{\lambda\sigma\nu}+(a+b)T_{\nu\sigma\lambda}+(-a+c)T_{\sigma\lambda\nu}+(-b-c)T_{\lambda\nu\sigma}=0. \tag{9b}$$

It is well known how every tensor can be decomposed as the sum of irreducible representations.[7] For the case of $T_{\mu\nu\lambda}$ there results:

$$T_{\mu\nu\lambda}=\overset{00}{T}_{\mu\nu\lambda}+\overset{01}{T}_{\mu\nu\lambda}+\overset{02}{T}_{\mu\nu\lambda}+\eta_{\mu\lambda}\overset{11}{T}_\nu+\eta_{\mu\nu}\overset{22}{T}_\lambda+\eta_{\mu\nu}a_\lambda+\eta_{\mu\lambda}a_\nu+\eta_{\nu\lambda}a_\mu+\epsilon_{\mu\nu\lambda\rho}b^\rho, \tag{10}$$

where $\overset{00}{T}_{\mu\nu\lambda}$ is completely symmetrical and of null trace, belonging to the representation $(\frac{3}{2},\frac{3}{2})$; $\overset{01}{T}_{\mu\nu\lambda}$ is antisymmetrical in $\mu\lambda$, of null trace, and it vanishes when it is completely antisymmetrized, belonging to the representation $(\frac{1}{2},\frac{3}{2})\oplus(\frac{3}{2},\frac{1}{2})$, $\overset{02}{T}_{\mu\nu\lambda}$ is antisymmetrical in $\mu\nu$, of null trace, and it vanishes when it is completely antisymmetrized, belonging to the representation $(\frac{1}{2},\frac{3}{2})\oplus(\frac{3}{2},\frac{1}{2})$. The terms which contain the 4-vector $\overset{11}{T}_\mu$, $\overset{22}{T}_\mu$, a_μ, and b^ρ belonging to the representations $(\frac{1}{2},\frac{1}{2})$, being $\epsilon_{\mu\nu\lambda\rho}$ completely antisymmetrical with $\epsilon^{0123}=+1$.

Introducing the decomposition (10) in Eq. (9b) we find that the verification of the latter imply total antisymmetry of $\overset{01}{T}_{\mu\nu\lambda}$ and $\overset{02}{T}_{\mu\nu\lambda}$, losing in this way the representation of the spin 2. We would like to point out that Eqs. (8) and (9) are satisfied for the irreducible parts which contain the 4-vector a_μ and the tensor $\overset{00}{T}_{\mu\nu\lambda}$ but they, by themselves, do not represent the spin 2. Besides, (8) and (9) imply the nullity of the terms in $\overset{11}{T}_\mu$ and $\overset{22}{T}_\nu$, and of the term $\epsilon_{\mu\nu\lambda\rho}b^\rho$. Therefore, it is demonstrated that it is not possible to write $R_{\mu\nu\lambda\sigma}$ as a function of a third rank tensor potential satisfying the covariance under the group C^*, according to Eq. (5).

We have arrived at the conclusion that the formulation of Ref. 4, although it keeps a close analogy with electromagnetic theory, it loses it when one develop the theory in terms of a potential. In the spin 2 and vanishing mass case there does not exist a potential whose transformation law under C^* leads us to the right transformation law (5) for the tensor $R_{\mu\nu\lambda\sigma}$, whereas the transformation law under C^* for the electromagnetic potential A_μ is consistent with the corresponding transformation law for the field strength tensor $F_{\mu\nu}$.

This fact does not allow the description of local interactions of the field with others, since, for doing so, introducing an appropriate potential with which to write the terms of the interaction is unavoidable.

*Fellow of the Consejo Nacional de Investigaciones Científicas y Técnicas.

[1]M. Fierz and W. Pauli, Proc. Roy. Soc. (London) A **173**, 211 (1939).
[2]E. Wigner, Ann. Math. **40**, 149 (1939).
[3]S. Weinberg, Phys. Rev. **138**, B988 (1965); **134**, B822 (1964).
[4]C. Coulter, Nuovo Cimento B **7**, 284 (1972).
[5]H. Kastrup, Phys. Rev. **150**, 1183 (1966), and references quoted therein.
[6]This potential was first used by S. Chang, Phys. Rev. **148**, 1259 (1966), for the quantization of the massive spin-2 field.
[7]See, for example, M. Castagnino, Mathematicae Notae **21**, 177 (1968).

Charge quantization and canonical quantization*

J. G. Miller

Department of Physics, University of Utah, Salt Lake City, Utah 84112
(Received 24 July 1975)

Dirac's charge quantization condition is derived by means of a canonical quantization procedure of an enlarged classical phase space in which the interaction constant is a dynamical variable. The charge quantization condition follows by imposing a superselection rule. The method avoids string singularities and does not depend on spherical symmetry. The charge quantization condition is due solely to the topology of the enlarged classical configuration space.

1. EQUATIONS OF MOTION

The nonrelativistic equations of motion for a particle of mass m and charge e moving in the field of a magnetic monopole of magnetic charge g fixed at the origin are given by the Lorentz force law

$$m\ddot{\mathbf{r}} = e\dot{\mathbf{r}} \times g\mathbf{r}/r^3. \tag{1}$$

(We use units throughout this paper in which $c = \hbar = 1$.) The equations can easily be integrated because of the existence of enough first integrals or constants of the motion. We find that

$$T = \tfrac{1}{2}mv^2, \quad \mathbf{J} = \mathbf{L} + \mathbf{j}, \quad \text{and} \quad L \tag{2}$$

are constants of the motion, where $\mathbf{L} = \mathbf{r} \times \mathbf{p}$, $\mathbf{p} = m\mathbf{v}$, is the orbital angular momentum and $\mathbf{j} = -eg\mathbf{r}/r$ is the angular momentum in the electromagnetic field due to the superposition of the electric field of the charged particle and the magnetic monopole field.[1] The total angular momentum is $\mathbf{J}$. The magnitude L of orbital angular momentum is a constant of the motion since $\mathbf{L}$ and $\mathbf{j}$ are orthogonal. Because $(\mathbf{r}/r) \cdot \mathbf{J} = -eg$, the motion of the charged particle is restricted to the surface of a right-circular cone.[2]

2. VECTOR POTENTIAL AND HAMILTON'S EQUATIONS

A Hamiltonian for the equations of motion (1) is given by

$$H = (1/2m)(\mathbf{p} - e\mathbf{A})^2, \tag{3}$$

where $\mathbf{p} = m\mathbf{v} + e\mathbf{A}$ is a canonical momentum and $\mathbf{A}$ is a vector potential for the magnetic monopole field $\mathbf{B} = g\mathbf{r}/r^3$. The problem is that a global vector potential $\mathbf{A}$ on $M = R^3 - \{0\} \approx R \times S^2$ does not exist. In the language of differential forms,

$$F = B_x\, dy \wedge dz + B_y\, dz \wedge dx + B_z\, dx \wedge dy = g \sin\theta\, d\theta \wedge d\phi \tag{4}$$

on M is closed but not exact. In order to quantize the motion of the charged particle, it has been customary to introduce a vector potential which is singular along a "string" or "strings" emanating from the origin and extending to infinity. For example, Schwinger[3] chose a vector potential given by

$$A = \frac{g}{2r}\left[\left(\frac{y}{r-z} - \frac{y}{r+z}\right)dx - \left(\frac{x}{r-z} - \frac{x}{r+z}\right)dy\right]$$

$$= -g\cos\theta\, d\phi\,. \tag{5}$$

There are two strings for this singular potential, the positive and negative z axis. The quantization proceeds by canonical quantization using this singular vector potential in (3). At this point the nature and meaning of the string singularities must be carefully examined. It is found that the string singularities correspond to strings of magnetic dipoles or, equivalently, semi-infinite solenoids,[4] and the singular field on these strings must be subtracted in order to obtain a theory of magnetic monopoles rather than a theory of semi-infinite solenoids.[5,6] A rigorous mathematical treatment of this problem has been given by Hurst,[7] who showed that a charge quantization condition ($\mu = eg$ is quantized) is obtained if and only if rotational invariance is imposed on the quantum mechanical Hamiltonian. Rather than creating problems at the outset in the classical theory by introducing a singular vector potential and then having to deal with the singularity in the quantum theory, it would be desirable to have a canonical quantization procedure that avoided string singularities altogether. The main purpose of this paper is to show that such a procedure exists.

3. NONCANONICAL SYMPLECTIC STRUCTURE ON T^*M

In the classical theory string singularities may be avoided in two ways. One way is to avoid the use of a vector potential. This can be done by choosing

$$\omega_F = \omega_M - e\tau_M^* F \tag{6}$$

as the symplectic 2-form on T^*M, where ω_M is the canonical symplectic 2-form on T^*M, i.e., $\omega_M = dx \wedge dp_x + dy \wedge dp_y + dz \wedge dp_z$ and τ_M: $T^*M \to M$ is the cotangent bundle projection, and by taking

$$\bar{H} = p^2/2m \tag{7}$$

as the Hamiltonian. The upper star on τ_M denotes the "pull-back" map. Here $\mathbf{p}$ is not a canonical momentum but is $m\mathbf{v}$. Hamilton's equations for this Hamiltonian and symplectic structure are equivalent to the Lorentz force law equations.[8,9] The Hamiltonian $\bar{H}$ is a constant of the motion and is the kinetic energy of the charged particle. The total angular momentum $\mathbf{J}$ is a constant of the motion since $\mathbf{J}$ commutes with $\bar{H}$, the Poisson bracket being computed with ω_F.[10] Also $\{J_x, J_y\} = J_z$ and cyclic permutations of x, y, z. In fact, $\mathbf{J}$ is the infinitesimal generator of the rotation group SO(3) acting

on T^*M and leaving ω_F and $\bar{H}$ invariant, i.e.,

$$\begin{aligned}(dJ_x)^{\#} &= y\partial_z - z\partial_y + p_y\partial_{p_z} - p_z\partial_{p_y},\\ (dJ_y)^{\#} &= z\partial_x - x\partial_z + p_z\partial_{p_x} - p_x\partial_{p_z},\\ (dJ_z)^{\#} &= x\partial_y - y\partial_x + p_x\partial_{p_y} - p_y\partial_{p_x},\end{aligned} \tag{8}$$

where $\#$ is the "index raising" operation[10] with respect to ω_F.

A quantization procedure based on an algebra of observables, where the commutation relations are obtained from the Poisson bracket structure in (T^*M, ω_F), has been given by Lipkin *et al.*[11] By a general procedure called prequantization,[12] Greub and Petry[13] are able to write a Schrödinger equation for a cross section in a certain complex line bundle. The prequantization condition is just Dirac's charge quantization condition. This approach is also described by Wu and Yang[14] in terms of a gauge field theory with structure group U(1). We now consider the second way in which string singularities in the classical theory may be avoided. This will lead to Dirac's charge quantization condition as a superselection rule. We will be dealing with principal fibre bundles rather than the associated vector bundles. For the connection between these two approaches, see Trautman.[15]

4. A FOUR-DIMENSIONAL CLASSICAL THEORY

The equations of motion for the charged particle can be obtained from the geodesic equations in a four-dimensional Riemannian (or pseudo-Riemannian) manifold (B, g_B), where B is a principal fibre bundle with base space M and one-dimensional structure group G and g_B is a Riemannian metric on B invariant under translations in the fibers. This is analogous to Kaluza's five-dimensional theory of relativity,[16] in which the equations of motion for the charged particle are obtained from the geodesic equations in a five-dimensional pseudo-Riemannian manifold.

Let $B = R^4 - \{0\} \approx R \times S^3$ and let X_i be left-invariant vector fields on the Lie group S^3, group manifold of SU(2), that satisfy

$$[X_1, X_2] = -X_3 \tag{9}$$

and cyclic permutations of 1,2,3. Let ω^i be the dual left-left-invariant 1-forms, i.e., $\omega^i(X_j) = \delta^i{}_j$. Then $d\omega^3 = \omega^1 \wedge \omega^2$ and cyclically. Define a Riemannian or pseudo-Riemannian metric g_B on B by

$$g_B = dr^2 + r^2[(\omega^1)^2 + (\omega^2)^2] + \kappa g^2(\omega^3)^2, \tag{10}$$

where r is the radial coordinate on $R^4 - \{0\}$ and κ is a nonzero constant. We note that this metric is the asymptotic limit of the NUT metric in case $\kappa = -1$ and $g = 2l$, where l is the NUT parameter. The NUT metric represents the gravitational field of a source with both mass and dual (or magnetic) mass.[17] Dowker[17] has pointed out the analogy between the work of Misner[18] on the NUT metric and the work of Hurst[7] on the magnetic monopole. We carry the analogy even further in considering the equations of motion. For our purposes κ is arbitrary but nonzero. We note that metric (10) has the same symmetry as the Lagrangian for the symmetric top.[19]

Let $G = S^1$ be the one-parameter subgroup generated by X_3 and let G act on S^3 and hence B by right multiplication. Then B is a principal fibre bundle with base space $B/G = M$ and structure group G and X_3 is a fundamental vector field.[20] $B(M, G)$ is the Cartesian product of R with the Hopf fibering of S^3 over S^2.[21] Now $L_{X_3} g_B = 0$, and so g_B is invariant under the action of G on B.

We follow Sniatycki and Tulczyjew's[9] formulation of Kaluza's theory. Define a 1-form ω on B by

$$\omega(X) = (\kappa g)^{-1} g_B(X_3, X), \tag{11}$$

where X is any vector field on B. We see that $\omega = g\omega^3$. Since ω is a contraction of $g_B \otimes X_3$, $L_{X_3}\omega = 0$. Thus ω is a connection 1-form[20] on the principal fibre bundle $B(M, G)$. The connection 1-form ω takes values in the Lie algebra $\mathfrak{G}$ of G, which is the set of real numbers. The vector fields X_1, X_2, and ∂_r are horizontal and X_3 is vertical. Since $L_{X_3} = i_{X_3} d + d i_{X_3}$ and $i_{X_3}\omega = \omega(X_3) = g$, $i_{X_3} d\omega = 0$. This implies that there exists a unique 2-form F on M such that

$$d\omega = \pi^* F, \tag{12}$$

where $\pi: B \to B/G$ is the fibre bundle projection.

In Euler angle coordinates defined by $q = q_0 + iq_1 + jq_2 + kq_3 = r \exp(k\phi/2) \exp(j\theta/2) \exp k\psi/2)$,

$$\begin{aligned}X_1 &= \sin\psi\partial_\theta - \csc\theta\cos\psi\partial_\phi + \cot\theta\cos\psi\partial_\psi,\\ X_2 &= \cos\psi\partial_\theta + \csc\theta\sin\psi\partial_\phi - \cot\theta\sin\psi\partial_\psi,\\ X_3 &= -\partial_\psi,\\ \omega^1 &= \sin\psi\, d\theta - \sin\theta\cos\psi\, d\phi,\\ \omega^2 &= \cos\psi\, d\theta + \sin\theta\sin\psi\, d\phi,\\ \omega^3 &= -d\psi - \cos\theta\, d\phi,\end{aligned} \tag{13}$$

and $\pi: R \times S^3 \to R \times S^2$ is given by $(r, \theta, \phi, \psi) \rightsquigarrow (r, \theta, \phi)$, where θ and ϕ become spherical coordinates on S^2. A calculation shows that F is given by (4). Although F is not exact on M, $\pi^* F$ is exact on B.

Let g_M be the Riemannian metric on M defined by

$$g_M(\pi_* hX, \pi_* hY) = g_B(hX, hY), \tag{14}$$

where hX and hY are the horizontal parts of the vector fields X and Y on B. A calculation shows that g_M is the Euclidean metric on M, which in spherical coordinates is given by

$$g_M = dr^2 + r^2(d\theta^2 + \sin^2\theta\, d\phi^2). \tag{15}$$

We decompose g_B as follows:

$$g_B(X, Y) = g_M(\pi_* hX, \pi_* hY) + \kappa\omega \otimes \omega(vX, vY), \tag{16}$$

where vX and vY denote the vertical parts of X and Y. A decomposition of the Taub–NUT metric similar to this was used by Hawking and Ellis[22] to explain its global properties. The geodesic equations in (B, g_B) are related to the Lorentz force law equations in (M, g_M, F). In fact, we will show that the set of all geodesics in (B, g_B) in some sense represent the set of all orbits in (M, g_M, F) for all values of the charge e. We should think of the fibre S^1 as being the dimension of charge.

Let $P = T^*B$ and let ω_B be the canonical symplectic 2-form on P. A Hamiltonian $H: P \to R$ for the geodesic

equations in (B,g_B) is given by

$$H(\alpha)=(1/2m)g_B^{-1}(\alpha,\alpha), \tag{17}$$

where g_B^{-1} is the contravariant metric tensor and $\alpha\in P$. The action of G on B lifts[10,23] to an action of G on P and this action leaves ω_B and H invariant. A theorem of Marsden and Weinstein[23,24] on the reduction of symplectic manifolds with symmetry now applies, and we obtain a reduced symplectic manifold in which this symmetry is divided out.

The momentum $P(X)$, where X is a vector field on B, is a function on $P=T^*B$ defined by

$$P(X)(\alpha)=\alpha(X(b)), \tag{18}$$

where $\alpha\in P$, $\tau_B(\alpha)=b$, and τ_B: $T^*B\to B$ is the cotangent bundle projection. A moment[23,24] Ψ: $P\to\mathfrak{G}^*$ for the action of G on P is given by

$$\Psi(\alpha)\cdot\xi=P(\xi X_3)(\alpha), \tag{19}$$

where $\xi\in\mathfrak{G}$ and $\alpha\in P$. The moment Ψ is equivariant with respect to the co-adjoint action of G on $\mathfrak{G}^*$ and every value in $\mathfrak{G}^*$ is a regular value of Ψ. We have that

$$\Psi^{-1}(\mu)=\{\alpha\in P\,|\,P(X_3)(\alpha)=\mu\}. \tag{20}$$

The isotropy group G_μ is the subgroup of G which leaves $\mu\in\mathfrak{G}^*$ fixed. Here $G_\mu=G$. The isotropy group G_μ leaves $\Psi^{-1}(\mu)$ invariant and, since it acts freely and properly on $\Psi^{-1}(\mu)$, $P_\mu=\Psi^{-1}(\mu)/G_\mu$ is a manifold.

The Marsden–Weinstein theorem states that there exists a unique symplectic 2-form ω_μ on P_μ such that

$$\pi_\mu^*\omega_\mu=i_\mu^*\omega_B, \tag{21}$$

where π_μ: $\Psi^{-1}(\mu)\to\Psi^{-1}(\mu)/G_\mu$ is the projection onto P_μ and i_μ: $\Psi^{-1}(\mu)\to P$ is the inclusion map. Since H is invariant under the action of G, the flow of $X_H=(dH)^\#$, the Hamiltonian vector field of H, induces a Hamiltonian flow on P_μ with Hamiltonian $\tilde{H}$ given by $\tilde{H}\circ\pi_\mu=H\circ i_\mu$. The reduced phase space is (P_μ,ω_μ) and the reduced Hamiltonian is $\tilde{H}$.

Let $\beta\in T_m^*M$ and let $b\in B$ such that $\pi(b)=m$. Define $\alpha\in T_b^*B$ by $\alpha(hX)=\beta(T\pi\circ hX)$ for any $X\in T_bB$ and $\alpha(X_3(b))=\mu$. $T\pi$ is the tangent of π.[10,23] Then $\alpha\in\Psi^{-1}(\mu)$. The mapping f: $T^*M\to P_\mu$ defined by $\beta\rightsquigarrow[\alpha]$ is well defined, where $[\alpha]$ is the corresponding equivalence class in P_μ. For $\mu=eg$, f:$(T^*M,\omega_F)\to(P_\mu,\omega_\mu)$ is a symplectic diffeomorphism, i.e., f is a diffeomorphism and $f^*\omega_\mu=\omega_F$, and $\bar{H}+e^2/2m\kappa=\tilde{H}\circ f$.

The power of the Marsden–Weinstein reduction theorem is that it is global. In Euler angle coordinates

$$H=\frac{1}{2m}\left[p_r^2+\frac{1}{r^2}\left(p_\theta^2+\frac{(p_\phi-\cos\theta\,p_\psi)^2}{\sin^2\theta}\right)+\frac{p_\psi^2}{\kappa g^2}\right] \tag{22}$$

and ψ is a cyclic coordinate. Thus p_ψ is a constant of the motion. If we set $p_\psi=-eg$ in H, we obtain the Hamiltonian (3) with vector potential (5) expressed in spherical coordinates plus the constant $e^2/2m\kappa$. Although the reduction can be carried out very simply in a coordinate system in which X_3 is a coordinate vector field, we obtain only a piece of the reduced phase space. From this coordinate point of view, H in (3) is $\bar{H}$ expressed in a symplectic chart, which does not exist globally since F is not exact.

There are two ways of interpreting (3) depending upon the choice of domain of H. If the singularities are excluded from the domain of H, then H is $\bar{H}$ expressed in a symplectic chart. If the domain of H is R^3, then H is the Hamiltonian for the charged particle in the field of one or more semi-infinite solenoids, depending upon the number of string singularities in **A**. The problem with implementing a canonical quantization of the symplectic structure (T^*R^3,ω_{R^3}), where ω_{R^3} is the canonical symplectic 2-form on T^*R^3, and taking H as the Hamiltonian is that we obtain a quantization of the motion of a charged particle in the field of one or more semi-infinite solenoids. It is claimed that if the resulting quantum mechanical Hamiltonian is spherically symmetric, it represents the Hamiltonian for the charged particle in the magnetic monopole field. It is found that this quantum mechanical Hamiltonian is spherically symmetric if and only if μ has discrete values.[7]

5. CANONICAL QUANTIZATION

We implement a canonical quantization in (T^*B,ω_B) by taking the Hilbert space $\mathcal{H}$ to be the set of complex-valued functions on B which are square integrable with respect to the invariant measure obtained from g_B. The quantum operator corresponding to a real-valued function f on B is multiplication by f and the operator corresponding to the momentum $P(X)$, where X is a vector field on B, is $P_X=-iX$.[23] The quantum operator corresponding to (17) is

$$H=-\frac{1}{2m}\Delta=\frac{1}{2m}\left(-\frac{1}{r}\,\partial_r(r\partial_r)+\frac{1}{r^2}(P_{X_1}^2+P_{X_2}^2)+\frac{1}{\kappa g^2}P_{X_3}^2\right), \tag{23}$$

where Δ is the Laplace–Beltrami operator on (B,g_B).

Before a physical interpretation is given, we must consider a reduction procedure just as in the classical case. We impose the superselection rule[25] corresponding to the operator P_{X_3}. The Hilbert space $\mathcal{H}$ is the direct sum of the mutually orthogonal eigenmanifolds of P_{X_3} and we project H onto an eigenmanifold $\mathcal{H}_\mu$, μ an eigenvalue of P_{X_3}. The projection H onto $\mathcal{H}_\mu$ is the Hamiltonian for the motion of the charged particle in the monopole field with interaction constant μ. The interaction constant μ has discrete values because the structure group G is compact. Since the Euler angle coordinate ψ is periodic with period 4π, we obtain Dirac's charge quantization condition[26]

$$eg=\tfrac{1}{2}n, \tag{24}$$

where n is an integer. The compactness of the structure group G follows from the requirement that the second Betti number of the principal fibre bundle $B(M,G)$ be zero so that π^*F is exact. We could have taken B to be $R\times P^3$ instead of $R\times S^3$, where P^3 is three-dimensional real projective space. P^3 is the group manifold of SO(3). In this case we obtain Schwinger's charge quantization condition[3]

$$eg=n. \tag{25}$$

P^3 is just the lens space[21] (2,1). We can take $B=R$

$\times B_k$ for the enlarged classical configuration space, where B_k is the lens space $(k,1)$, since the second Betti number of B vanishes for each positive integer k. The charge quantization condition is then

$$eg = \tfrac{1}{2}kn, \tag{26}$$

where k is a fixed positive integer and n is the quantum number. These charge quantization conditions have been obtained by Usachev[27] by taking a vector potential with k singular strings. However, the number of strings does not enter into our formulation because $\pi_k^* F$ is exact on $B = R \times B_k$, where π_k is the bundle projection π_k: $R \times B_k \to R \times S^2$. There are only two more principal fibre bundles B with base space M and one-dimensional fibre G and they are both trivial. They are $M \times S^1$ and $M \times R$. The second Betti number of both is one and $\pi^* F$ is closed but not exact on both these bundles, where π is the projection onto the first factor.

We have rederived Hurst's results from a different point of view. The Hamiltonian (23) after the superselection rule is imposed differs from Hurst's only by a constant. It is identical to Hurst's Hamiltonian in the limit as κ approaches infinity. The P_{X_i} correspond to Hurst's $\tilde{J}$'s[7] or Peshkin's K's.[28]

The charge quantization condition does not depend on the spherical symmetry of the monopole field. Indeed, let $\tilde{F} = F + \bar{F}$, where $\bar{F}$ is any exact 2-form on M and let $\tilde{\omega} = \omega + \pi^* \sigma$, where $\bar{F} = d\sigma$. For example, $\bar{F}$ could be a uniform magnetic field. Then $d\tilde{\omega} = \pi^* \tilde{F}$. Define a Riemannian or pseudo-Riemannian metric $\tilde{g}_B$ on B, invariant under translations in the fibers, by

$$\tilde{g}_B(X, Y) = g_M(\pi_* hX, \pi_* hY) + \kappa \tilde{\omega} \otimes \tilde{\omega}(vX, vY), \tag{27}$$

where κ is a nonzero constant and hX and vX are the horizontal and vertical parts of the vector field X with respect to the connection 1-form $\tilde{\omega}$ on $B(M,G)$. Then

$$\tilde{g}_B = dr^2 + r^2[(\omega^1)^2 + (\omega^2)^2] + \kappa(\tilde{\omega})^2. \tag{28}$$

We can show exactly as before that the geodesic equations in $(B, \tilde{g}_B)$ after a reduction procedure are the Lorentz force law equations in $(M, g_M, \tilde{F})$. The same charge quantization condition is obtained as before by imposing the superselection rule corresponding to the operator P_{X_3}. In general, $\tilde{F}$ has no symmetry and the only symmetry that $\tilde{g}_B$ possesses is that generated by X_3.

ACKNOWLEDGMENTS

I wish to thank the referee for a helpful comment that improved the paper and Ronald Stern for discussions on topology.

*Work supported in part by NSF Grants GP-38953 and GP-43718X.

[1]M. Fierz, Helv. Phys. Acta 17, 27 (1944).
[2]H. Poincaré, Compt. Rend. 123, 530 (1896).
[3]J. Schwinger, Phys. Rev. 144, 1087 (1966).
[4]P.O. Grönblom, Z. Phys. 98, 283 (1935).
[5]P. Jordan, Ann. Physik 32, 66 (1938).
[6]G. Wentzel, Progr. Theor. Phys. Suppl. No. 37—38, 163 (1966).
[7]C.A. Hurst, Ann. Phys. (N.Y.) 50, 51 (1968).
[8]J.-M. Souriau, *Structure des systèmes dynamiques* (Dunod, Paris, 1970).
[9]J. Sniatycki and W.H. Tulczyjew, Ann. Inst. H. Poincaré 15, 177 (1971).
[10]R. Abraham, *Foundations of Mechanics* (Benjamin, New York, 1967).
[11]H.J. Lipkin, W.I. Weisberger, and M. Peshkin, Ann. Phys. (N.Y.) 53, 203 (1969).
[12]J. Sniatycki, J. Math. Phys. 15, 619 (1974).
[13]W. Greub and H.-R. Petry, J. Math. Phys. 16, 1347 (1975).
[14]T.T. Wu and C.N. Yang, "Concept of nonintegrable phase factors and global formulation of gauge fields," submitted to Phys. Rev.
[15]A. Trautman, Repts. Math. Phys. 1, 29 (1970).
[16]E. Leibowitz and N. Rosen, General Relativity Gravitation 4, 449 (1973).
[17]J.S. Dowker, General Relativity Gravitation 5, 603 (1974).
[18]C.W. Misner, J. Math. Phys. 4, 924 (1963).
[19]R. Hermann, *Differential Geometry and the Calculus of Variations* (Academic, New York, 1968), p. 225.
[20]S. Kobayashi and K. Nomizu, *Foundations of Differential Geometry* (Interscience, New York, 1963), Vol. 1.
[21]N. Steenrod, *The Topology of Fibre Bundles* (Princeton U.P., Princeton, N.J., 1951).
[22]S.W. Hawking and G.F.R. Ellis, *The Large Scale Structure of Space-Time* (Cambridge U.P., Cambridge, 1973), p. 175.
[23]J. Marsden, *Applications of Global Analysis in Mathematical Physics* (Publish or Perish, Boston, 1974).
[24]J. Marsden and A. Weinstein, Repts. Math. Phys. 5, 121 (1974).
[25]R.F. Streater and A.S. Wightman, *PCT, Spin and Statistics, and All That* (Benjamin, New York, 1964), p. 5.
[26]P.A.M. Dirac, Proc. Roy. Soc. (London) A 133, 60 (1931).
[27]Yu. D. Usachev, a paper presented at the All-Union Conference of the Universities of the Soviet Union held at Uzhgorod, October, 1968.
[28]M. Peshkin, Ann. Phys. (N.Y.) 66, 542 (1971).

On the dynamics of particles in a bounded region: A measure theoretical approach

C. Marchioro* and A. Pellegrinotti*

Istituto Matematico, Università di Roma, Rome, Italy

E. Presutti†

Mathematical Department, University of Stanford, Stanford, California

M. Pulvirenti

Istituto Matematico, Università di Camerino, Camerino (Mc), Italy
(Received 20 February 1975; revised manuscript received 3 July 1975)

An existence theorem is proven for the solution of the differential equations of motion of a finite number of particles moving in a bounded piecewise regular region and mutually interacting via C^1 forces. It is shown that the elastic reflection laws uniquely determine a Lebesgue measurable flow solution of the differential equations of motion (with elastic boundary conditions). The Lebesgue measure is invariant so that an extension of the Liouville theorem to non-Hamiltonian flows is obtained. A natural representation of the time evolution is given as a flow upward from a base under a "ceiling" function.

1. INTRODUCTION

In recent papers it has been shown that it is possible to reduce the problem of the existence of dynamics for an infinite system to the problem of n particles moving in a bounded region Λ.[1-4] This kind of result is of interest in a statistical mechanics framework, see Refs. 5–8. As far as we know, in spite of its apparent elementariness, there does not seem to be any exhaustive study on the argument.

The usual approach is to consider n particles, interacting via regular forces (see D 2.2), in the open region Λ. By well-known existence theorems, dynamics is defined until a particle reaches the boundaries $\partial\Lambda$, of Λ. The motion is then extended to later times by elastic reflection prescriptions.

There are pathologies connected with this procedure. Problems arise from collisions with zero normal velocity: If the particle would naturally escape from the region Λ, in the absence of the walls, the simple elastic reflection laws would not be sufficient to specify the further motion of the particle. It is also possible that during the motion a particle has infinitely many collisions with the boundaries in a finite time. Further, if one assumes $\partial\Lambda$ not to be completely smooth, see D 2.1, a particle may reach $\partial\Lambda$ in a singular point and then the elastic reflection prescriptions would not make sense. Finally, the presence of the infinite forces representing the action of the walls could undermine the invariance of the Lebesgue measure $\nu\ [=d(q)_n d(p)_n]$, the classical proof of Liouville's theorem applies to Hamiltonian flows.

We will prove that the initial data in which the above pathologies may be present are in some sense exceptional: We can exclude a set of null ν-measure and in the remaining of the phase space a ν-invariant global flow is constructed. The particles have finitely many collisions in a bounded interval of time, and never hit $\partial\Lambda$ in its singular part or with zero normal velocity.

The main steps in the proof of the above results are the following.

The first one is to imbed the finite system (Γ, ν), $\Gamma = \Lambda^n \times \mathbb{R}^{\nu n}$ into the unbounded dynamical system $(\Gamma^\infty, \nu^\infty, S(t))$, $\Gamma^\infty = \mathbb{R}^{\nu n} \times \mathbb{R}^{\nu n}$, ν^∞ is the Lebesgue measure on $\mathbb{R}^{2\nu n}$ and $S(t)$ is the flow determined by the differential equations of motion. We then consider the S-trajectories in Γ^∞, in those intervals of time τ^+ determined by the entrance and exit of some particle from Λ. The main point is to show that the union of these "cut" trajectories is a ν-measurable set in Γ, and that a. e. w. the above intervals of time (for which the trajectories are in Γ) are strictly positive. Then time evolution on this set is naturally represented as a special flow upward starting from a base, with "ceiling" function τ^+ and base measure μ^+, the measure projected from ν on the base along the S-flow.[9,10] To obtain a global evolution, it is then necessary to determine a transformation R of the base, which is meant to connect a point in the ceiling to a point in the base in a μ^+-preserving way. The upward flow and R therefore define a transformation T on the base, under which μ^+ is invariant. By general arguments (Poincaré's recurrence theorem, see Ref. 11), this is sufficient to ensure global evolutions in Γ independently of the characteristics of T and μ^+.

In our particular case the main problems are essentially three.

The first is to prove measurability for the above sets, and this is obtained with topological arguments. The second is the proof that the elastic reflection prescriptions determine a μ^+-invariant transformation of the base; this task is accomplished by explicit knowledge of μ^+. The third is the proof that τ^+ is μ^+-strictly positive, obtained by using continuity properties of the motion.

In Sec. 2 we give definitions and notations and then develop the above arguments in some detail. The technical estimates are reported in the form of theorems and their proofs are given in Sec. 3. In Sec. 4 some conclusions of this paper are drawn.

2. DEFINITIONS, RESULTS AND OUTLINE OF THE PROOF

In the first five definitions below, we specify the hypotheses on the system we treat.

D 2.1 *Space of configurations*: By Λ, we denote an open bounded set in $\mathbb{R}^\nu$. We assume that its boundary $\partial\Lambda$ is piecewise smooth: closure of a finite union of surfaces with continuous normal derivative. The particles we study move in $\Lambda \cup \partial\Lambda$.

D 2.2 *Interactions*: Let $U: \mathbb{R}^{\nu n} \to \mathbb{R}$ be a C^2 bounded below function representing the potential energy of our system of n particles so that

$$F_i = \text{force on the } i\text{th particle} = -\frac{\partial}{\partial q_i} U[(q)_n], \qquad (2.1)$$

$$(q)_n = q_1, \ldots, q_n.$$

D 2.3 *Phase spaces*: By Γ, we denote the phase space of our system:

$$\Gamma = \Lambda^n \times \mathbb{R}^{\nu n}, \quad \Gamma = \{x \in \mathbb{R}^{2\nu n} \mid x = ((q)_n(p)_n),\ q_i \in \Lambda,\ p_i \in \mathbb{R}^\nu \text{ for } i = 1, \ldots, n\}. \qquad (2.2)$$

It will be convenient to think of Γ as a subspace of the unbounded phase space, $\Gamma^\infty = \mathbb{R}^{\nu n} \times \mathbb{R}^{\nu n}$, of n particles moving in the whole space. Γ^∞ is equipped with the topology of $\mathbb{R}^{2\nu n}$, Γ is an open set in Γ^∞ and sometimes it will be thought of as a topological space with the induced topology of Γ^∞. The actual motion will take place in the closure $\overline{\Gamma}$ of Γ.

D 2.4 *Measures*: We denote by ν^∞, the completed Lebesgue measure on Γ^∞ by ν its restriction to Γ. Both ν^∞ and ν are σ-finite measures. By ν_E^∞ (ν_E), we will denote the relativization of ν^∞ (ν), to the measurable sets with energy less than E. Notice that by D 2.2 the sets of configurations with energy less than E is open and therefore ν^∞ (ν)-measurable.

D 2.5 *Time evolution in* Γ^∞: For every real t we denote by $S_t: \Gamma^\infty \to \Gamma^\infty$, the time evolution in Γ^∞ determined by the interaction in D 2.2:

$$\frac{d}{dt} p_i(t) = \frac{\partial}{\partial q_i} U[(q)_n], \quad m\frac{d}{dt} q_i(t) = p_i(t), \qquad (2.3a)$$

$$p_i(0) = p_i, \qquad q_i(0) = q_i, \qquad i = 1, \ldots, n. \qquad (2.3b)$$

In the sequel we shall consider the map $\bar{\psi}: \Gamma^\infty \times \mathbb{R} \to \Gamma^\infty$, defined as

$$\bar{\psi}(x, t) = S_t x. \qquad (2.4)$$

$\bar{\psi}$ is continuous if $\Gamma^\infty \times \mathbb{R}$ is equipped with the product topology. In particular we will use the following stronger property of S_t:

P_1 For every $x \in \Gamma^\infty$, $\epsilon > 0$, $T > 0$ there exists a neighborhood V_x of x such that

$$|S_t x - S_t y| < \epsilon, \quad y \in V_x, \quad |t| \leq T, \qquad (2.5)$$

(where the norm in Eq. 2.5 is that of $\mathbb{R}^{2\nu n}$).

In order to represent (part of) the flow S_t, see the outline of the proof in the Introduction (as a flow upward from a base, it is convenient to use explicit notations for the sets in which some particle is in $\partial\Lambda$), definitions D 2.6, and D 2.7 below.

D 2.6 *Decomposition of the boundaries*: Let n be the inward unit vector normal to $\partial\Lambda$, defined in $\partial\Lambda - \partial\Lambda_s$, $\partial\Lambda_s$ is the singular part of $\partial\Lambda$, see D 2.1. We will consider the following sets in Γ^∞:

$$\mathcal{J}^\pm = \{x = (q)_n(p)_n \mid \text{for } i = 1, \ldots, n \quad \text{either } q_i \in \Lambda \text{ or } q_i \in \partial\Lambda - \partial\Lambda_s \text{ and } p_i \mathrm{n} \gtrless 0.\ \text{There exists } j \text{ s.t. } q_j \dot{\in} \partial\Lambda\},$$

$$\mathcal{J}^0 = \{x = (q)_n(p)_n \mid \text{for } i = 1, \ldots, n, \quad q_i \in \Lambda \cup \partial\Lambda \text{ and there exists } j \text{ s.t. } q_j \in \partial\Lambda_s, \text{ or } q_j \in \partial\Lambda \text{ and } p_j \mathrm{n} = 0\}.$$

$\mathcal{J} = \mathcal{J}^0 \cup \mathcal{J}^+ \cup \mathcal{J}^-$ will be considered equipped with the topology induced by Γ^∞. In this topology both $\mathcal{J}^+$ and $\mathcal{J}^-$ are open sets.

D 2.7 *Time return to* $\partial\Lambda$: For every $x \in \Gamma$ there exists an open nonzero maximal connected time interval $I(x)$ s.t. for $t \in I(x)$, $S_t x \in \Gamma$. For $x \in \mathcal{J}^+$ ($\mathcal{J}^-$) correspondingly there exists a maximal positive (negative) function $\tau^+(\tau^-)$, s.t. $S_t x \in \Gamma$ for $0 < t < \tau^+(x)$ [$\tau^-(x) < t < 0$]. In this time interval the motion Eqs. (2.3) make sense as evolution in Γ; our task will be to show how to extend them for times larger [smaller] than $\tau^+(x)$ [$\tau^-(x)$] by taking into account the elastic reflection laws.

The first step in the construction of dynamics is to consider the subset in Γ of configurations evoluted from $\mathcal{J}^+$. More precisely let

$$M^+ = \{y \in \mathcal{J}^+ \times \mathbb{R}^+ \mid y = (x, t), \quad 0 < t < \tau^+(x), \quad x \in \mathcal{J}^+\}, \qquad (2.6a)$$

$$\Gamma^+ = \bar{\psi}(M^+) \subset \Gamma, \qquad (2.6b)$$

and assume analogous definitions for M^- and Γ^-.

In Theorems 2.1—2.3 we state properties of $\mathcal{J}^\pm$, $\tau^\pm$, and $\Gamma^\pm$, even if for the sake of brevity (both in the theorems and in their proofs); we only refer to the "+" part.

Theorem 2.1: M^+ is an open set in $\mathcal{J}^+ \times \mathbb{R}^+$ (equipped with the product topology), and Γ^+ is an open set in Γ. The restriction of $\bar{\psi}$ to M^+ is denoted by ψ and it is a one to one bicontinuous mapping of M^+ onto Γ^+. The function $\tau^+(x)$ defined on Γ^+ is lower semicontinuous.

We pointed out the above topological properties of M^+ and Γ^+ in order to have measure theoretical information. The purpose is to project ν on $\mathcal{J}^+$ along the flow S_t. This will be obtained in the theorem stated after the following definitions.

D 2.8 *Measures on* $\mathcal{J}$: Let σ be the completed orthogonal projection of ν on $\mathcal{J}$. In particular, $\sigma^\pm$ is its restriction to $\mathcal{J}^\pm$. Since $\mathcal{J}^0$ has null σ measure and so does the set of configuration which has more than one particle in $\partial\Lambda$, the following function:

$$\pi_{\mathrm{n}}: \mathcal{J} \to \mathbb{R}, \quad \pi_{\mathrm{n}}(x) = p_i \mathrm{n}, \text{ where } p_i \text{ is the momentum of the only particle in } \partial\Lambda, \qquad (2.7)$$

is σ-almost everywhere defined. We will often consider the σ-finite measures $\mu^\pm = \pi_{\mathrm{n}} \circ \sigma^\pm$ on $\mathcal{J}^\pm$.

D 2.9 *Lebesgue measure on* $\mathbb{R}^+$: By λ we denote the Lebesgue measure on $\mathbb{R}^+$.

Theorem 2.2: Let $\mu^+\times\lambda$ be the product measure on $\mathcal{J}^+\times\mathbb{R}^+$ and $\mu^+\times\lambda\,|M^+$ its restriction to M^+. Let $\nu^\infty|\Gamma^+ = \nu|\Gamma^+$ be the restriction of ν^∞ to Γ^+, then:

(i) $\{M^+, \mu^+\times\lambda\,|M^+\}$ is isomorphic to $\{\Gamma^+, \nu^\infty|\Gamma^+\}$,

(ii) τ^+ is μ^+-measurable and the set $\{x\in\mathcal{J}^+\,|\,\tau^+(x)=\infty\}$ has null μ^+ measure,

(iii) $\tau^+(x)^{-1}$ is μ^+ measurable and if $A\subset\mathcal{J}^+$ is any μ^+ measurable set

$$\int_{\mathcal{J}^+}\mu^+(dx)\,\chi(A,x)=\int_{\Gamma^+}\nu[d\psi(x,t)]\,\chi(A,x)\,\tau^+(x)^{-1},$$

where $\chi(A,x)$ denotes the value in x of the characteristic function of the set A.

We want to show that the above theorem defines μ^+-essentially a map from $\mathcal{J}^+$ to $\mathcal{J}^-$. This and the successive arguments will finally be collected in Theorem 2.4 stated below. We define μ^+-a.e.w. a map S^+,

$$S^+\colon\ \mathcal{J}^+\to\mathcal{J}^-\cup\mathcal{J}^0\quad S^+x=S_{\tau^+(x)}x,\quad x\in\mathcal{J}^+. \tag{2.8}$$

Since by Theorem 2.2 (ii) $\tau^+(x)$ is μ^+-essentially finite, the only problem is to show that the set of those $x\in\mathcal{J}^+$, s.t. $S^+x\in\mathcal{J}^0$, are μ^+ negligible. In the following theorem we give estimates on the measure of these sets. We first establish some notations.

D 2.10 *Special sets*: Analogous to Eq. (2.6), we pose

$$M^0=\{y\in\mathcal{J}^0\times\mathbb{R}\,|\,y=(x,t),\quad x\in\mathcal{J}^0,\quad S_{t'}x\in\bar\Gamma \text{ for } 0\le t'\le t \text{ or } t\le t'\le 0\},$$

$$\Gamma^0=\bar\psi(M^0).$$

Therefore, Γ^0 contains the set of all the configuration in $\bar\Gamma$ which in their natural motion (without any previous collision) hit the boundary in its singular part $\partial\Lambda_s$, or reach $\partial\Lambda$ with zero normal velocity. We will have to show that Γ^0 is ν-negligible and also that the set of configurations in Γ, which after some collision enter into Γ^0, is also ν-negligible. We will first show this by studying the sets

$$\mathcal{N}^+=\{x\in\mathcal{J}^+\,|\,S^+x\in\mathcal{J}^0\},$$

$$\Gamma_{\mathcal{N}^+}=\{y\in\Gamma^+\,|\,y=S_tx\quad x\in\mathcal{N}^+\quad t<\tau^+(x)\}.$$

Theorem 2.3: (i) The set Γ^0 has null ν measure. In particular $\nu(\Gamma_{\mathcal{N}^+})=0$ (ii) The set $\mathcal{N}^+$ has μ^+-null measure.

(ii) of Theorem 2.3 proves indeed that S^+ is μ^+-essentially defined from $\mathcal{J}^+$ to $\mathcal{J}^-$.

Proof that $S^+\mathcal{J}^+$ *has full* μ^-*-measure*: In analogy to D 2.7, we considered the negative time return $\tau^-(x)$ to $\partial\Lambda$ defined from $\mathcal{J}^-$ to $\mathcal{J}$. Analogous theorems to Thm. 2.1–2.3 hold so that a μ^--a.e.w. defined map S^- from $\mathcal{J}^-$ to $\mathcal{J}^+\cup\mathcal{J}^0$ is considered. The set

$$\mathcal{N}^-=\{x\in\mathcal{J}^-\,|\,S^-x\in\mathcal{J}^0\}$$

has μ^- null measure.

We have

$$[\mathcal{J}^- - \mathcal{N}^- - \{x\in\mathcal{J}^-\,|\,\tau^-(x)=\infty\}]\subset S^+\mathcal{J}^+. \tag{2.9}$$

In fact, if y is in the lhs of Eq. 2.9, it means that $S^-y\in\mathcal{J}^+$ and by the same Eqs. of motion, $S^+S^-y=y$. Therefore,

$$\mu^-[(S^+\mathcal{J}^+)^{\text{compl}}]\le\mu^-(\mathcal{N}^-)+\mu^-[\{x\in\mathcal{J}^-\,|\,\tau^-(x)=\infty\}]=0.$$

As a result we have proven that S^+ is a (modulo zero) invertible map from $(\mathcal{J}^+,\mu^+)$ to $(\mathcal{J}^-,\mu^-)$.

Further, from the explicit form of μ^+ given in (iii) of Theorem 2.3, and the analogous form for μ^- (A is μ^+ measurable), we have

$$\begin{aligned}\mu^+(A)&=\int_{\Gamma^+}\nu[d\psi(x,t)]\,\chi(A,x)\,\tau^+(x)^{-1}\\&=\int_{\Gamma^-}\nu[d\psi(S^+x,t-\tau^+(x))]\,\chi(S^+A,S^+x)\,|\tau^-(S^+x)^{-1}|\\&=\mu^-(S^+A).\end{aligned}\tag{2.10}$$

Therefore, the measure spaces $(\mathcal{J}^+,\mu^+)$ and $(\mathcal{J}^-,\mu^-)$ are isomorphic with the isomorphism given by S^+. By explicit construction D 2.8, $(\mathcal{J}^-,\mu^-)$ is also isomorphic to $(\mathcal{J}^+,\mu^+)$ by means of the transformation R (elastic reflection),

$$R(q)_n(p)_n=(q)_n(p)_n',$$

where $(p)_n'$ is equal to $(p)_n$ except for the ith particle, which is in $\partial\Lambda$ and whose normal momentum has opposite sign.

As a consequence, the transformation T

$$T=RS^+\colon\ \mathcal{J}^+\to\mathcal{J}^+, \tag{2.11}$$

is μ^+a.e.w. defined and is an automorphism of $(\mathcal{J}^+,\mu^+)$ onto itself.

It still remains to prove that the set

$$\mathcal{B}=\{x\in\mathcal{J}^+\,|\,\textstyle\sum_0^\infty{}_n\,\tau^+(T^nx)<\infty\} \tag{2.12}$$

has null μ^+ measure. This is proven by general arguments (Poincaré's recurrence theorem, see for instance Ref. 11)

Proof that $\mu^+(\mathcal{B})=0$: Let

$$\mathcal{A}_a^E=\{x\in\mathcal{J}^+\,|\,\tau^+(x)>a,\quad E(x)<E\}.$$

Since

$$\mu^+[\{x\in\mathcal{J}^+\,|\,\tau^+(x)=0\}]=0,$$

by countable additivity of μ^+ we have

$$\mu^+(\mathcal{B})=\lim_{a\to c}\lim_{E\to\infty}\mu^+(\mathcal{A}_a^E\mathcal{B}).$$

By Poincaré's recurrence theorem ($\mu^+(\mathcal{A}_a^E\mathcal{B})=0$): almost all x in $\mathcal{A}_a^E\mathcal{B}$ appear infinitely many times in $\mathcal{A}_a^E\mathcal{B}$; but since this is impossible by the very definition of $\mathcal{B}$ and $\mathcal{A}_a^E$, it must be that $\mu^+(\mathcal{A}_a^E\mathcal{B})=0$.

A corollary of the above result is that the set

$$\Gamma_{\mathcal{B}}^+=\{y\in\Gamma^+\,|\,y=(x,t),\quad x\in\mathcal{B}\text{ or }T^nx\in\mathcal{N}^+\text{ for }n\ge0,\quad t<\tau^+(x)\},$$

has null ν measure. In fact, the base of $\Gamma_{\mathcal{B}}^+$ has null μ^+ measure (proven before) and by using Theorem 2.2 the desired result is obtained.

We collect the above arguments in the theorem below which is therefore proven.

Theorem 2.4: The map S^+ defined in Eq. 2.8 is an isomorphism between the two Lebesgue spaces $(\mathcal{J}^+, \mu^+)$ and $(\mathcal{J}^-, \mu^-)$. Therefore, the map $T = RS^+$ where R is the elastic reflection map, see Eq. 2.11, is an automorphism of $(\mathcal{J}^+, \mu^+)$ onto itself.

The time evolution S_t restricted to $\overline{\Gamma}^+ = \Gamma^+ \cup \mathcal{J}^+ \cup S^+\mathcal{J}^+$, and extended by the elastic reflection laws, defines ν a. e. w. a flow T_t in the following sense:

(i) $\overline{\Gamma}^+$ is T_t-invariant (ν modulo zero),

(ii) the set of configurations which have infinitely many collisions in a finite time, or hit $\partial\Lambda$ in its singular part, or reach at some time $\partial\Lambda$ with zero normal velocity, has null measure.

The dynamical system $(\overline{\Gamma}^+, \nu|\overline{\Gamma}^+, T_t)$, is represented as a special flow with ceiling function τ^+ under a base $\mathcal{J}^+$ with base transformation T.

Theorem 2.4 does not yet complete the study of the configurations which have particles colliding on the walls $\partial\Lambda$. It still remains to consider the configurations in Γ^0 after some collisions. Theorem 2.3 (i) proves that $\nu(\Gamma^0) = 0$ while the configurations of Γ^+ entering into Γ^0 belong to $\mathcal{N}\Gamma^+$, see D 2.10, and therefore have null ν measure.

The configurations in Γ whose particles never suffer collision dynamics are directly defined via the flow S_t therefore, the required existence theorem is proven; the above results are stated in Theorem 2.5 below.

Theorem 2.5: The phase space $\overline{\Gamma}$ is ν-essentially the disjoint union of the two invariant sets $\overline{\Gamma}^+$ and $\overline{\Gamma}^f$. $\overline{\Gamma}^+$ is described in Theorem 2.4 and Γ^f is the set of configurations whose particles never suffer collisions. As a conclusion, the set $(\overline{\Gamma}, \nu, T_t)$ is a dynamical flow: $T_t x$ represents the unique solution of the Eqs. of motion 2.3 with elastic reflections on the boundaries. The number of collisions is finite for every bounded interval of time and no particle hits $\partial\Lambda$ in its singular part or with zero normal velocity.

3. PROOFS

In this section we give the proofs of Theorems 2.1—2.3.

Proof of Theorem 2.1: The proof of Theorem 2.1 is carried through in a sequence of steps stated below as propositions; some of them are so obvious they do not require a proof.

Proposition 1: $\mathcal{J}^+$ is open in $\mathcal{J}$.

Proposition 2: Γ^+ is open in Γ.

Proof: Let $y \in \Gamma^+$, $y = \psi(x, t)$, and $x \in \mathcal{J}^+$, $t < \tau^+(x)$. Then there exists $T > t$ such that

$$S_{-t'} y \in \overline{\Gamma} \quad \text{for } t < t' \leq T.$$

As a consequence, there exists $\epsilon > 0$ so that:

$$\text{(i)} \quad \inf_{0 \leq t' \leq T} |S_{-t'} y - z| < \epsilon \Rightarrow z \in \mathcal{J}^0 \text{ for } z \in \Gamma^\infty,$$

(ii) $S_{-T} y$ has distance from $\mathcal{J}$ which is strictly larger than ϵ.

The existence of such an ϵ is ensured by the fact that $\mathcal{J}^0$ is closed. We now apply P_1 of D 2.5: There exists V_y open in Γ^0 such that

$$\text{for } z \in V_y, \quad |S_{-t'} y - S_{-t'} z| < \epsilon, \quad 0 \leq t' \leq T.$$

By property (i) $S_{-t'} z$ is never in $\mathcal{J}^0$, and by (ii), and the continuity of the motion, it crosses $\mathcal{J}^+$ so that it belongs to Γ^+.

Proposition 3: $\bar{\psi}$ is continuous.

Proposition 4: M^+ is open in $\mathcal{J}^+ \times \mathbb{R}^+$ and $\tau^+(x)$ is lower semi continuous.

Proof: Since $M^+ = \bar{\psi}^{-1}(\Gamma^+)$ with Γ^+ open and $\bar{\psi}$ continuous, M^+ is open in $\mathcal{J}^+ \times \mathbb{R}^+$. $\tau^+(x)$ is l. s. c.: we will show that $x_n \to x$ in $\mathcal{J}^+$, and $\lim_n \inf \tau^+(x_n) < \tau^+(x)$ leads to a contradiction. Let

$$\liminf_{n \to \infty} \tau^+(x_n) = \lim_{k \to \infty} \tau^+(x_{n_k}) = T < T' < \tau^+(x), \tag{3.1a}$$

$$\lim_{k \to \infty} x_{n_k} = x \tag{3.1b}$$

Then (x_{n_k}, T') for $k > k_0$ does not belong to M^+ because k_0 is so chosen that

$$\text{for } k > k_0, \quad T' > \tau^+(x_{n_k}).$$

Since $\lim_{k \to \infty} (x_{n_k}, T') = (x, T') \in M^+$, this is absurd because M^+ was proven to be open.

Proposition 5: ψ^{-1} is continuous.

Proof: Let $y_n \to y$ in Γ^+, with $\psi(x_n, t_n) = y_n$, $\psi(x, t) = y$. We will show that $x_n \to x$ and $t_n \to t$. We construct T and ϵ as in the proof of Proposition 2. Then we choose subsequences $\{n_k\}$ and $\{n_j\}$ such that:

$$\bar{t} = \limsup_{n \to \infty} t_n = \lim_{k \to \infty} t_{n_k} < T,$$

$$\underline{t} = \lim_n \inf t_n = \lim_{j \to \infty} t_{n_j},$$

$$\bar{x} = \lim_{k \to \infty} x_{n_k},$$

$$\underline{x} = \lim_{j \to \infty} x_{n_j}$$

We have:

$$\lim_{k \to \infty} y_{n_k} = \lim_{k \to \infty} \psi(x_{n_k}, t_{n_k}) = \psi(\bar{x}, \bar{t}),$$

$$\lim_{k \to \infty} y_{n_k} = y = \psi(x, t),$$

and since ψ has an inverse, by its definition and the Eqs. of motion, Eq. 2.3, it follows that $\bar{x} = x$ and $\bar{t} = t$. Analogously, we have that $\underline{x} = x$ and $\underline{t} = t$.

Proof of Theorem 2.2: (i) First we notice that by Theorem 2.1, (Γ^+, ν) is isomorphic to (M^+, ν_ψ) where $\nu_\psi = \nu \circ \psi$. We then define $\mu^+ \times \lambda$ on $\mathcal{J}^+ \times \mathbb{R}^+$, and consider its restriction to M^+, which is an open set by Theorem 2.1 and therefore $\mu^+ \times \lambda$ measurable. We compare $\mu^+ \times \lambda$ and ν_ψ on a class $\mathcal{T}$ of Borel sets (tubes) which σ-generates the whole σ-algebra of measurable sets. We say that $B \in \mathcal{T}$, if B is the open set of the form

$$B = \{y \in M^+ \mid y = (x, t), x \in \mathcal{A} \text{ open in } \mathcal{J}^+, 0 < t_1 < t < t_2 < \tau^+(x)\}.$$

It is easy to see that $\mathcal{T}$ σ-generates the algebra of Borel sets. We can then define

$$\mu_t^+(\mathcal{A}) = (t)^{-1} \circ \nu_\psi[\mathcal{A} \times (0, t)],$$

which is t-continuous. Furthermore, by Liouville's theorem on the unbounded system $(\Gamma^\infty, \nu, S_t)$,

$$\mu^+_{t/k}(\mathcal{A}) = \mu^+_t(\mathcal{A}) \quad \text{for every integer } k.$$

Hence, μ^+_t is constant on a dense set and therefore does not depend on t, $\mu^+_t = \hat{\mu}^+$. The proof that $\hat{\mu}^+$ equals μ^+ can be obtained by simply computing the Jacobian of the transformation

$$(x_0, t) \to S_t x_0.$$

This gives the normal momentum π_n, which enters in the definition of μ^+, see D 2.8. Therefore, (i) of Theorem 2.2 is proved.

(ii) By Theorem 2.1, $\tau^+(x)$ is lower semicontinuous and therefore measurable w.r.t. the Borel measure μ^+. Let

$$\mathcal{G}^\infty_E = \{x \in \mathcal{J}^+ \,|\, \tau^+(x) = \infty, \quad E(x) = \text{energy of } x < E\}.$$

Then ν_E and μ^+_E are finite measures so that

$$\mu^+_E(\mathcal{G}^\infty_E) = T^{-1}\, \nu_\varphi[\mathcal{G}^\infty_E \times (0,T)] \leq T^{-1}\, \nu(\{x \in \Gamma \,|\, E(x) < E\}) \underset{T\to\infty}{\to} 0,$$

where (i) of Theorem 2.2 has been used. This proves (ii) of Theorem 2.2.

(iii) This is a consequence of the above proven measurability of $\tau^+(x)$. By use of Fubini's theorem we have

$$\begin{aligned}
&\int_{\Gamma^+} \nu[d\psi(x,t)]\, \chi(\mathcal{A}, x)\, \tau^+(x)^{-1} \\
&\quad = \int_{\mathcal{J}^+ \times \mathbb{R}^+} \mu^+(dx) \times \lambda(dt)\, \chi(\mathcal{A}, x)\, \tau^+(x)^{-1} \chi[M^+, (x,t)] \\
&\quad = \int_{\mathcal{J}^+} \mu^+(dx)\, \chi(\mathcal{A}, x)\, \tau^+(x)^{-1} \int_{\mathbb{R}^+} \lambda(dt)\, \chi[M^+, (x,t)] \\
&\quad = \int_{\mathcal{J}^+} \mu^+(dx)\, \chi(\mathcal{A}, x)\, \tau^+(x)^{-1} \int_0^{\tau^+(x)} dt = \int_{\mathcal{J}^+} \mu^+(dx)\, \chi(\mathcal{A}, x).
\end{aligned}$$

Proof of Theorem 2.3: The main point in the proof of Theorem 2.3 is the following argument, introduced as a separate lemma.

Lemma 3.1: Let σ be the surface measure on $\mathcal{J}$, B a σ-measurable set, $Q = \cup_{0\,t} S_t B$. We have

$$\nu^\infty_{\text{ext}}(Q_T) \leq w(B)\, |T|\, \sigma(\bar{B}),$$

where ν^∞_{ext} is the outer measure associated to ν^∞, $w(B)$ is the supremum of the moduli of the momenta of the particles in the configurations in $\bar{B}$, $\bar{B}$ is the closure of B in $\mathcal{J}$.

Proof: ν^∞_{ext} is a S-invariant outer measure on Γ^∞. In fact ν^∞_{ext} is defined by the following:

$$\nu^\infty_{\text{ext}}(B) = \inf_{\{A_n\}} \textstyle\sum_n \nu^\infty(A_n), \quad B \subset \cup A_n,\ A_n \text{ is } \nu^\infty \text{ measurable.}$$

Therefore,

$$\nu^\infty_{\text{ext}}(B) = \inf_{\{A_n\}} \sum \nu^\infty(A_n) = \inf_{\{S_t A_n\}} \sum \nu^\infty(S_t A_n) \geq \nu^\infty_{\text{ext}}(S_t B),$$

$$\nu^\infty_{\text{ext}}(S_t B) = \inf_{\{C_n\}} \sum \nu^\infty(C_n) = \inf_{\{S_{-t} C_n\}} \sum \nu^\infty(S_{-t} C_n) \geq \nu^\infty_{\text{ext}}(B).$$

We therefore have

$$\nu^\infty_{\text{ext}}(Q_T) \leq m\, \nu^\infty_{\text{ext}}(Q_{T/m}), \qquad m \in \mathbb{Z}^+. \tag{3.2}$$

To obtain an estimate on the r.h.s. of Eq. 3.2 we consider the following ν^∞-measurable set containing $Q_{T/m}$: for every configuration in B take the neighborhood for which every particle is within

$$[w(B) + F\,|T|/m]\,|T|/m,$$

with momentum within $F\,|T|/m$, where F is the maximal force that a particle can experience. This set is open and therefore ν^∞-measurable, it obviously contains $Q_{T/m}$; if its projection on $\mathcal{J}$ is denoted by $C(B, T/m)$, we have from Eq. 3.2

$$\begin{aligned}
\nu^\infty_{\text{ext}}(Q_T) &\leq m\, \sigma[C(B, T/m)]\, (w(B) + F\,|T|/m)\, |T|/m \\
&\underset{m\to\infty}{\to} \sigma(\bar{B})\, w(B)\, |T|.
\end{aligned}$$

This proves Lemma 3.1.

Proof of (i) *of Theorem* 2.3: Let

$$M^0_{T,E} = \{y \in M^0 \,|\, y = (x,t), \quad |t| \leq T, \quad E(x) \leq E\}.$$

We have by Lemma 3.1 that

$$\nu_{\text{ext}}[\bar{\psi} M^0_{T,E}] \leq \nu^\infty_{\text{ext}}[\textstyle\bigcup_{0\,|t|}^{T} S_t\, \mathcal{J}^0_E] \leq 2\, w(\mathcal{J}^0_E)\, T\, \sigma(\mathcal{J}^0) = 0,$$

so that (i) is proven when we let T and E diverge.

Proof of (ii) *of Theorem* 2.3: Let

$$\mathcal{M}_{E,T} = \{x \in \mathcal{N}^+ \,|\, E(x) < E,\ \tau^+(x) > T\}.$$

Then, by the previous estimates

$$\nu_{\text{ext}}[\textstyle\bigcup_{0\,t}^{T} S_t\, \mathcal{M}_{E,T}] \leq \nu(\Gamma_{\mathcal{N}^+}) = 0.$$

Therefore, by Theorem 2.2,

$$T\, \mu^+(\mathcal{M}_{E,T}) = \nu[\textstyle\bigcup_{0\,t}^{T} S_t\, \mathcal{M}_{E,T}] = 0,$$

Letting T^{-1} and E diverge, the proof is complete.

4. CONCLUSIONS

Here we consider the following two problems: Are the pathological configurations also negligible with respect to the microcanonical measures? Does the technique we used in this paper apply to collisions between particles as in the case of hard-core interactions?

In the remainder of this section we will give some sketchy arguments on a possible way to treat the above problems.

We proved in Theorems 2.4 and 2.5 that the catastrophic configurations we are dealing with are in a Lebesgue null set. However, they may not be negligible w.r.t. the Lebesgue measure projected on surfaces of constant energy, the microcanonical measures.

To prove that this is not the case we could proceed as in Secs. 2 and 3. Since the energy surfaces $\sum(E)$, are closed we directly derive from Theorems 2.1 and 2.2 their analogous just by considering intersections with $\sum(E)$. The arguments used in the proof of Lemma 3.1 carry through in this new case, so Theorem 2.3 can also be proved. Since the results of Theorem 2.4 and 2.5 are consequences of the validity of Theorems 2.1–2.3, we may obtain the existence theorem for finite volume dynamics in a set of full measure w.r.t. all the microcanonical measures associated to the system.

Dynamics of finitely many hard-cores in a bounded region: In this case the elastic reflection law has to be extended to collisions between particles: It is of particular interest for the very same definition of dynamics to show that multiple (more than two particles) collisions are present in a set of null Lebesgue measure.

The first problem which arises in this case is to find out the "good" infinite system in which to imbed our "singular" hard-core finite one. This will be the only point we sketch about this problem.

The idea is that the infinite regularized dynamical system, still denoted by $(\Gamma^\infty, \nu^\infty, S_t)$, is in fact determined by n fictitious point particles moving in $\mathbb{R}^\nu$. Their interactions equal the actual ones (between the hard cores) whenever the positions of the fictitious point particles are consistent with the hard core restriction and which are extended to regular interaction in the remaining configurations.

The base of our flow will then be extended in the phase space to cover situations in which the particles reach the hard-core distance. Again at this point, topological considerations would have to be used in order to prove the measure estimates needed in the extensions of the theorems of Sec. 2.

ACKNOWLEDGMENTS

We are grateful to Giovanni Gallavotti for interesting discussions, to the referee of the J. Math. Phys. for very useful comments on a previous version of this paper and to Dan Rudolph for suggesting deep simplifications in the proof of Theorem 2.4. One of us (E. P.) acknowledges very kind hospitality at the Mathematical Department of Stanford University.

*Research partially supported by CNR.

†On leave of absence from Istituto Matematico, Universita' dell'Aquila, L'Aquila, Italy. Research partially supported by a CNR fellowship, Posit. 204.530.

[1]Ya. G. Sinai, Vestn. Mosk. Univ. **1**, 152 (1974).
[2]C. Marchioro, A. Pellegrinotti, and E. Presutti, Lett. Nuovo Cimento **11**, 606 (1974).
[3]C. Marchioro, A. Pellegrinotti, and E. Presutti, Commun. Math. Phys. **40**, 175 (1975).
[4]O. E. Lanford III. Lecture notes, University of Seattle (1974).
[5]O. E. Lanford III, Commun. Math. Phys. **9**, 176 (1968).
[6]O. E. Lanford III, Commun. Math. Phys. **11**, 257 (1969).
[7]Ya. G. Sinai, Sov. Theor. Math. Phys. **12**, 487 (1973).
[8]E. Presutti, M. Pulvirenti, and B. Tirozzi, Nota interna n. 544, Istituto Fisico, Univ. Roma, 1974.
[9]W. Ambrose, Ann. Math. **42**, 3 (1941).
[10]W. Ambrose and S. Kakutani, Duke Math. J. **9**, 25 (1942).
[11]P.R. Halmos, *Lecture on Ergodic Theory* (Chelsea, New York, 1958).

Conditioning of states

Platon C. Deliyannis

Department of Mathematics, Illinois Institute of Technology, Chicago, Illinois 60616
(Received 22 August 1975)

A system of axioms for the state space of a quantum system is proposed which, together with the concept of conditioning a state by the occurrence of an event, leads to the construction of the standard orthomodular events system.

INTRODUCTION

In the present paper we are exploiting the point of view advocated by several writers (Refs. 1 to 7), that the state space rather than the event (or observable) system is the natural underlying concept for the foundations of physics. Our main goal is to show how, by means of reasonable hypotheses, one can derive a meaningful concept of event and impose on the set of all events the standard structure of an orthomodular partially ordered set.

The state space we consider consists exclusively of pure states, because we feel that unavoidably one has to assume the principle of mixture: any state is a mixture (possible in several ways) of pure states. In such a setup the fundamental question is: what is the probability of (random) transition from one state to another? So we are led to postulate a number $\langle m \mid n \rangle$ corresponding to every pair of states m, n in our collection $\mathcal{P}$ of all pure states; randomness imposes symmetry on this functional. Note that in the classical situation we have $\langle m \mid n \rangle = 0$ for all distinct m, n.

We now come to the concept of event. Our interpretation (not quite original, to be sure) is that we can detect the occurrence of events by watching how the various states change. These changes are of no "duration"—time is not supposed to enter the picture, neither causal relationships. We are thus led to consider an event A completely determined if we know how each state of the system will change when this event A occurs. So we define an event A as a map from $\mathcal{P}$ to $\mathcal{P}$, which we shall write as $m \to m_{:A}$; the exact meaning of this being that when the system is in the state m, the event A occurs iff we detect a change from m to $m_{:A}$. A technical point arises now: how can we incorporate the case of an event A not being possible to occur in a state m? One possibility is to assign to each event a "domain", i.e., the set of states in which it is possible for A to occur, and assign meaning to the symbol $m_{:A}$ only in case m is in this domain. Another, is to introduce a hypothetical (or impossible) state, and incorporate the case of A not possibly occurring in the state m by saying that $m_{:A}$ is this fictitious state. We shall adopt the second alternative as it is technically simpler, but also physically suggestive.

According to our interpretation, it is clear that $\langle m \mid m_{:A} \rangle$ should be the probability $m(A)$ of A occurring in the state m. We shall assume this explicitly. One of the results we shall establish that makes our whole approach consistent, is that if the system is in some state m and we consider all states in which A occurs with certainty ($m_{:A}$ is one of them!) then $m_{:A}$ is characterized as that state which maximizes the transition probability $\langle m \mid n \rangle$.

A basic hypothesis which we shall adopt is what we call the subspace principle. Once an event, say A, has occurred, our state space is transformed from the original $\mathcal{P}$ to the set $\mathcal{P}_A$ of all states in which A occurs with certainty, according to our basic interpretation. But this cannot change the basic structure of the state space; hence we assume that any property of $\mathcal{P}$ has to be valid for all $\mathcal{P}_A$ also.

Before we enter into the details let us briefly mention how the structure of $\mathcal{P}$ can impose structure on the events. First, consider the meaning of implication. To say that A implies B means that B is bound to occur wherever A does; this we can reformulate as $m_{:A}(B) = 1$ for all states m. The basic properties follow, one of which is that the set of states in which A occurs with certainty determines A completely. The concept of opposite or complementary event is somewhat more involved. Given an event A we say that its opposite A' occurs with certainty in a state m iff A cannot occur in this state. We can see that A' is uniquely defined, but its existence must be postulated. Furthermore, it appears that in general the role A and A' is not symmetric, i.e., $(A')'$ need not be A; it is not hard to show that $A' = ((A')')'$ without extra hypotheses, and the possibility of a "Browerian" structure on the events obtainable from simple physical hypotheses appears to be of some interest. It turns out that $A = (A')'$ is true iff we assume that A is certain in a state m iff A' is impossible in m. This, as well as other standard properties of our class of events, follows from a general assumption we shall make on the behavior of maximal systems of states which are pairwise exclusive (i.e., have zero transition probabilities): if $\{m_i\}$ is such a set of states, then $\sum \langle m \mid m_i \rangle = 1$ for any state m. Note that an "orthogonal" system of states is just a classical system; thus, the above axiom means that given a maximal (i.e., exhaustive) classical subsystem of states, the probabilities of transition from any state to one of them are also exhaustive. Technically this means that any maximal set of mutually exclusive atomic events must have the certain event as a logical disjunction.

These are all the required hypotheses. We now proceed with the details.

1. EVENTS

The fundamental object in our construction is a set $\mathcal{P}$, which represents the pure states of the system, and

a functional $P\times P\to[0,1]$, to represent the probabilities of spontaneous transitions from one pure state to another. For m, $n\in P$ we shall write $\langle m\,|\,n\rangle$ for this probability. The first hypothesis we make is the following:

Axiom 1: (a) $0\leq\langle m\,|\,n\rangle\leq 1$,

(b) $\langle m\,|\,n\rangle=1$ iff $m=n$,

(c) $\langle m\,|\,n\rangle=\langle n\,|\,m\rangle$.

Part (a) is obviously essential. Part (b) reflects the noncausality of our transitions: if the system keeps oscillating between m, n, how can we distinguish between them? Part (c) also reflects noncausality, but in a more subtle way.

We shall call two pure states m, n *orthogonal* if $\langle m\,|\,n\rangle=0$; we shall write this as $m\perp n$.

It is technically convenient to introduce a fictitious state θ such that $\langle\theta\,|\,\theta\rangle=\langle m\,|\,\theta\rangle=\langle\theta\,|\,m\rangle=0$ for all $m\in P$. We shall write P_0 for $P\cup\{\theta\}$.

As discussed in the Introduction, we shall view an event as a transformation on our pure state space. It pays to allow the state θ to enter the game. Thus an event, say A, shall associate to each $m\in P_0$ another element of P_0 which we shall write as $m_{:A}$ and call "the state m conditioned by the occurence of A." It is understood that occurrence of A while the system is in the state m is tantamount to the system's switching over to $m_{:A}$. Thus the number $\langle m\,|\,m_{:A}\rangle$ shall be interpreted as the probability $m(A)$ of occurrence of A while the system is in the state m.

Formally we pose the definition as follows:

Definition 1. An *event* is a map $A: P_0\to P_0$ such that:

(i) A is idempotent, i.e., $m_{:A:A}=m_{:A}$, while $\theta_{:A}=\theta$,

(ii) $\langle m\,|\,m_{:A}\rangle=0$ implies either m or $m_{:A}=\theta$,

(iii) for any $m, n\in P_0$ we have $\langle m\,|\,n_{:A}\rangle=\langle m\,|\,m_{:A}\rangle\times\langle m_{:A}\,|\,n_{:A}\rangle$.

Keeping in mind the interpretation of $\langle m\,|\,m_{:A}\rangle$, we see that the set of all $m_{:A}$ which are not θ is precisely the set of states in which A occurs with certainty; hence reoccurrence of A should not change $m_{:A}$, i.e., A should be idempotent. The requirement $\theta_{:A}=\theta$ together with (ii) simply says that $m_{:A}=\theta$ iff it is impossible for A to occur in the state m. Requirement (iii) is strong but natural, and stresses the spontaneity of transitions by stating that the probability of switching from the arbitrary state m to a state p ($\equiv n_{:A}$) in which A occurs with certainty is the product (independence!) of the probability of A occurring, and the probability of the subsequent switching of $m_{:A}$ to p.

Definition 2: The set $D_A=\{m\,|\,m_{:A}\neq\theta\}$ is the *domain* of A and the set $R_A=\{m\in P\,|\,m=m_{:A}\}$ is the *range* of A. As mentioned above, D_A is the set of all states in which A is possible and R_A the set of those in which A is certain. Evidently $R_A\subseteq D_A$.

A symmetrized (equivalent) version of (iii) shall be useful in what follows. Take any m, $n\in P_0$ and consider $\langle m\,|\,m_{:A}\rangle\langle m_{:A}\,|\,n\rangle$: the second factor is then equal to $\langle m_{:A}\,|\,n_{:A}\rangle\langle n_{:A}\,|\,n\rangle$ by (iii), so the whole thing is $\langle m\,|\,m_{:A}\rangle\langle m_{:A}\,|\,n_{:A}\rangle\langle n_{:A}\,|\,n\rangle$. But the first two factors give just $\langle m\,|\,n_{:A}\rangle$ again by (iii). So we finally get:

$$\langle m\,|\,m_{:A}\rangle\langle m_{:A}\,|\,n\rangle=\langle n\,|\,n_{:A}\rangle\langle n_{:A}\,|\,m\rangle,$$

which we shall call the "symmetry" relation.

Example 1: (classical model) We take any set P and let $\langle m\,|\,n\rangle$ be the Kronecker symbol δ_{mn}. Our axiom is trivially verified. The events are in a 1:1 correspondence with the subsets of P. This is because by (ii) we have either $m_{:A}=\theta$ or $m_{:A}=m$, and we can associate with each A the set R_A; vice versa, for any subset S of P the map

$$m\to\begin{cases}\theta & m\notin S\\ m & m\in S\end{cases}$$

defines an event.

Example 2: (nonclassical model) We take P to be the set of all rays in some Hilbert space and represent them by unit vectors; write m_ϕ, m_ψ, etc. for the states represented by ϕ, ψ, etc. We set $\langle m_\phi\,|\,m_\psi\rangle=|(\phi\,|\,\psi)|^2$, where $(\phi\,|\,\psi)$ is the inner product. It is convenient in this instance to represent θ by the zero vector. Each projection operator determines an event. To see this let P be a projection and let $m_{:P}$ be represented by $P\psi$ (normalized in case it is $\neq 0$); properties (i), (ii), (iii) follow easily. It is not quite clear to the writer whether each event can be so represented by a projection. This shall be the case, however, if we assume the hypotheses to be introduced later.

We shall close this section with a result connecting D_A to R_A, offering some insight into our structure. For any $S\subseteq P$ we write $S^\perp$ for the set $\{m\in P\,|\,m\perp n$ for all $n\in S\}$, and cS for the set complement $\{m\in P\,|\,m\notin S\}$.

Proposition 1: For any $A\in\mathcal{L}$ we have $D_A=c(R_A^\perp)$.

Proof: We have $m\in cD_A$ iff $m_{:A}=\theta$, which means $\langle m_{:A}\,|\,n\rangle=0$ for all $n\in P$, which is equivalent to $\langle m_{:A}\,|\,n_{:A}\rangle\langle n_{:A}\,|\,n\rangle=0$. Thus we have by symmetry, $m\in cD_A$ iff $\langle n\,|\,n_{:A}\rangle\langle n_{:A}\,|\,m\rangle=0$ for all $n\in P$, and in particular $m\perp n$ for all $n\in R_A$. So $cD_A\subseteq R_A^\perp$. Conversely, if $m\in R_A$, we have $\langle m\,|\,n_{:A}\rangle=0$ for all $n\in P$, and reversing the above argument we get $m_{:A}=\theta$, or $m\in cD_A$.

2. PARTIAL ORDERING

We shall now impose structure on $\mathcal{L}$. First we consider the concept of implication.

Definition 3: The event B implies the event A iff $m_{:B}(A)=1$ for all states $m\in D_B$. We write this as $B\leq A$.

Note that this condition simply means that if B has occurred, then A is certain to occur also.

First we formulate this relation in various useful ways.

Proposition 1: We have $B\leq A$ iff $R_B\subseteq R_A$.

Proof: If $B\leq A$ and $m\in R_B$, then $m=m_{:B}$. Hence $m(A)=m_{:B}(A)$, or $m(A)=1$; thus $\langle m\,|\,m_{:A}\rangle=1$ and $m=m_{:A}$, or $m\in R_A$. Conversely, if $R_B\subseteq R_A$, take any $m\in D_B$; then $\theta\neq m_{:B}\in R_B$, so $m_{:B}\in R_A$ and $m_{:B}(A)=1$.

Proposition 2: We have $B\leq A$ iff $m_{:B}=m_{:A:B}$ and $m(B)=m(A)m_{:A}(B)$.

Proof: Suppose $B \leq A$. If $m \notin \mathcal{D}_B$, then $m_{:B} = \theta$, hence $m_{:B:A} = \theta$ also, or $m_{:B} = m_{:B:A}$. If $m \in \mathcal{D}_B$ then $m_{:B}(A) = 1$, i.e., $\langle m_{:B} \mid m_{:B:A} \rangle = 1$, or again $m_{:B} = m_{:B:A}$. Now take any n and use symmetry repeatedly to obtain: $n(B)\langle n_{:B} \mid m \rangle = \langle n \mid n_{:B} \rangle \langle n_{:B} \mid m \rangle = \langle m \mid m_{:B} \rangle \langle m_{:B} \mid n \rangle$ $= \langle m \mid m_{:B} \rangle \langle m_{:B} \mid m_{:B:A} \rangle \langle m_{:B} \mid n \rangle = \langle m \mid m_{:B} \rangle \langle m_{:B} \mid m_{:B:A} \rangle$ $\times \langle m_{:B:A} \mid n \rangle = \langle m \mid m_{:B} \rangle \langle n \mid n_{:A} \rangle \langle n_{:A} \mid m_{:B} \rangle = \langle n \mid n_{:A} \rangle$ $\times \langle m \mid m_{:B} \rangle \langle m_{:B} \mid n_{:A} \rangle = \langle n \mid n_{:A} \rangle \langle n_{:A} \mid n_{:A:B} \rangle \langle n_{:A:B} \mid m \rangle$ $= n(A)n_{:A}(B) < n_{:A:B} \mid m \rangle$. That is, for all m, n we have $n(B)\langle n_{:B} \mid m \rangle = n(A)n_{:A}(B)\langle n_{:A:B} \mid m \rangle$. In case $n_{:B} = \theta$, we have $n_{:A:B} = \theta$ also; because $n_{:A:B} \neq \theta$ implies $n_{:A} \neq \theta$, and so all three terms on the right will be $\neq 0$ for $m = n_{:A:B}$. Thus $n(B) = 0$, $n_{:A}(B) = 0$ and the result holds. In case $n_{:B} \neq \theta$ we have $n(B) \neq 0$; but we also have $n_{:A:B} \neq \theta$, for otherwise the right hand side is 0 for all m, while the left is not. By the next lemma we then get $n(B) = n(A)n_{:A}(B)$, and since these are $\neq 0$, we also have $n_{:B} = n_{:A:B}$. Now for the converse, take $m \in \mathcal{R}_B$ so that $m_{:B} = 1$. Then $m_{:A:B} = m$, hence $m_{:A} \in \mathcal{R}_B$, or $m_{:A}(B) = 1$ which gives $m(B) = m(A)$. Since $m(B) = 1$ we get $m(A) = 1$ or $m \in \mathcal{R}_A$.

Lemma 1: If $a_1\langle m_1 \mid m \rangle = a_2\langle m_2 \mid m \rangle$ (where m_1, $m_2 \neq \theta$) for all $m \in \mathcal{P}$, then $a_1 = a_2$; if they are $\neq 0$ we also get $m_1 = m_2$.

Proof: Take $m = m_1$ to obtain $a_1 = a_2\langle m_2 \mid m_1 \rangle \leq a_2$, and similarly $a_2 \leq a_1$. So if a_1, $a_2 \neq 0$, we get $\langle m_1 \mid m \rangle = \langle m_2 \mid m \rangle$ for all m, and in particular for $m = m_1$ we obtain $\langle m_2 \mid m_1 \rangle = 1$, or $m_1 = m_2$.

Theorem 1: The relation of implication is a partial order on $\mathcal{L}$.

Proof: Using Proposition 1, we obtain $A \leq A$ and transitivity at once. So assume $B \leq A$ and $A \leq B$. Then $m_{:A} = m_{:A:B}$, $m_{:B} = m_{:B:A}$, and $m_{:B} = m_{:A:B}$. Thus, $m_{:A} = m_{:B}$ for all m, i.e., $A = B$.

Corollary: The set $\mathcal{R}_A$ determines A completely.

The exact way of obtaining A out of $\mathcal{R}_A$ can be obtained as follows. This interpretation of the state $m_{:A}$ is quite natural and appears to justify the present approach.

Theorem 2: Given A and $m \in \mathcal{D}_A$, then $m_{:A}$ is the unique $n \in \mathcal{R}_A$ which maximizes $\langle m \mid n \rangle$.

Proof: Since $n \in \mathcal{R}_A$, we have $\langle m \mid m_{:A} \rangle \langle m_{:A} \mid n \rangle = \langle m \mid n \rangle$; thus $\langle m \mid n \rangle \leq \langle m \mid m_{:A} \rangle$ for all $n \in \mathcal{R}_A$, and $m_{:A}$ maximizes $\langle m \mid n \rangle$. Now if $\langle m \mid m_{:A} \rangle = \langle m \mid n \rangle$ for some $n \in \mathcal{R}_A$, then again we have $\langle m \mid m_{:A} \rangle \langle m_{:A} \mid n \rangle = \langle m \mid m_{:A} \rangle$, and since $\langle m \mid m_{:A} \rangle \neq 0$ (since $m \in \mathcal{D}_A$) we obtain $\langle m_{:A} \mid n \rangle = 1$, or $m_{:A} = n$ and uniqueness is established.

We shall now introduce several important events, and incidentally establish that $\mathcal{L} \neq \phi$.

Proposition 3: The maps

$$\begin{cases} O\colon\ m \to \theta \quad \forall\ m \in \mathcal{P}_0 \\ I\colon\ m \to m \quad \forall\ m \in \mathcal{R}_0 \end{cases}$$

are events, and $O \leq A \leq I$ for all $A \in \mathcal{L}$.

No proof is really needed.

Proposition 4: For each $m \in \mathcal{P}$ the map

$$A_m : n \to \begin{cases} m & \text{if } \langle n \mid m \rangle \neq 0 \\ \theta & \text{if } \langle n \mid m \rangle = 0 \end{cases}$$

is an event in $\mathcal{L}$.

Proof: Clearly θ is mapped to θ, and to finish (i) of definition 1 we need idempotency. Consider an n not orthogonal to m; then it is mapped by A_m to m, which again is not $\perp$ to m, so that we have it in this case. If $n \perp m$, then it is mapped to θ which again is mapped to θ. Thus A_m is idempotent. To verify (ii) let $m_1 = n_{:A_m}$ and note that if n, $m_1 \neq \theta$, then $m_1 = n$ while $\langle n \mid m \rangle \neq 0$, which is precisely the desired conclusion. For (iii) we want $\langle p \mid n_{:A_m} \rangle = \langle p \mid p_{:A_m} \rangle \langle p_{:A_m} \mid n_{:A_m} \rangle$, which holds for $n_{:A_m} = \theta$; so consider $n_{:A_m} \neq \theta$, which means that it is m, and the desired relation is $\langle p \mid m \rangle = \langle p \mid p_{:A_m} \rangle \langle p_{:A_m} \mid m \rangle$. Since $p_{:A_m}$ is θ or m according to whether $p \perp m$ or not, this evidently holds.

Even though the next result is obvious by Proposition 1, we state it as a theorem because of its importance.

Theorem: The events A_m are the atoms of $\mathcal{L}$.

Proof: Clearly each A_m is an atom, because $\mathcal{R}_{A_m}$ has no proper subsets, being just $\{m\}$. Conversely, if $A \neq O$, then $\mathcal{R}_A \neq \phi$; but if $\mathcal{R}_A$ contains m_1, m_2, $O \leq A_{m_1} \leq A$, $A_{m_1} \neq O$, $A_{m_1} \neq A$ and A is not an atom. So any atom A has $\mathcal{R}_A = \{m\}$ for some $m \in \mathcal{L}$, i.e., $A = A_m$.

Remark: O, I and the atoms A_m may very well be the only elements of $\mathcal{L}$, unless we assume more than Axiom 1. For example, let $\mathcal{P} = \{a, b, c\}$ with $\langle m \mid n \rangle = \frac{1}{2}$ for $m \neq n$. If $\mathcal{R}_A = \{a, b\}$ say, then A maps a to a, b to b and c to one of these, say b (it makes no difference). Then the requirement of symmetry yields $\langle a \mid a_{:A} \rangle \langle a_{:A} \mid c \rangle$ $= \langle c \mid c_{:A} \rangle \langle c_{:A} \mid a \rangle$ or $\langle a \mid c \rangle = \langle c \mid b \rangle \langle b \mid c \rangle$, which is absurd.

To close the section, we shall state our next axiom which shall be used in the next section. Consider two orthogonal states m, n and the corresponding atoms A_m, A_n; occurrence of A_m means that our system is in the state m, hence subsequent (spontaneous) occurrence of A_n is not possible, since $m(A_n) = \langle m \mid n \rangle = 0$. Thus the events A_m, A_n are very strongly mutually exclusive. Now, keeping this in mind, we consider a maximal orthogonal family $\{m_i\}$ of states; this means that no event A_n exclusive of each and every A_{m_i} exists. It is thus reasonable to expect that $\sum m(A_{m_i}) = 1$ for any state $m \in \mathcal{P}$. This we shall assume, explicitly, since it is formally independent of Axiom 1.

Axiom 2: Given any maximal orthogonal family $\{m_i\}$ of states, we have $\sum \langle m \mid m_i \rangle = 1$ for any $m \in \mathcal{P}$.

Note that the three-element example mentioned previously as well as examples 1 and 2 of Sec. 1, all satisfy Axiom 2.

3. THE SUBSPACE PRINCIPLE

The hypothesis we shall introduce now is motivated by the observation that once an event A has occurred, the various states of the system shall switch to their images under the map A (according to our basic interpretation), and so the state system $\mathcal{P}$ can be replaced by $\mathcal{R}_A$; in other words, unless something else occurs, the only states which make sense to look at are those

in which A occurs with certainty, simply because A has actually occurred! But this means that $\mathcal{R}_A$ must satisfy the same conditions as $\mathcal{P}$ does. There is no problem as far as Axiom 1 is concerned, since the restriction of $\langle\ |\ \rangle$ will evidently satisfy the required conditions. So let us write $\mathcal{L}_A$ for the set of all events based on $\mathcal{R}_A$, and use the same θ as for $\mathcal{L}$. We shall establish a natural 1:1 correspondence between the events of $\mathcal{L}_A$ and the events of $\mathcal{L}$ which are $\leq A$.

Given any $B \leq A$ we can easily obtain an element of $\mathcal{L}_A$ by simply restricting the map B to $\mathcal{R}_A$; since $\mathcal{R}_B \subseteq \mathcal{R}_A$, the values of the restriction are contained in $\mathcal{R}_A$ and evidently idempotency follows. Writing $\hat{B}$ for this restriction, we note that $\mathcal{D}_{\hat{B}}$, being $\{m \in \mathcal{R}_A \mid m_{:\hat{B}} \neq \theta\}$ is simply $\mathcal{D}_B \cap \mathcal{R}_A$. So $m \in \mathcal{D}_{\hat{B}}$ implies $\langle m \mid m_{:B}\rangle \neq 0$, and since $m_{:B} = m_{:\hat{B}}$, we see that condition (ii) holds. Requirement (iii) is naturally valid in $\mathcal{R}_A$ again because $m_{:\hat{B}} = m_{:B}$. Note that $\mathcal{R}_B$ and $\mathcal{R}_{\hat{B}}$ are identical.

The converse requires a little more work. We shall show that given any event $C \in \mathcal{L}_A$, we can find an event $\check{C} \in \mathcal{L}$, $\check{C} \leq A$, whose restriction to $\mathcal{R}_A$ is precisely C; since $\mathcal{R}_{\check{C}} = \mathcal{R}_C$ we see that such a $\check{C}$ is unique.

The definition of $\check{C}$ is obvious; it will just be the composition of the two maps A and C which we shall denote by $C \circ A$: $\check{C} = C \circ A$. It is evident that $\check{C}$ restricted to $\mathcal{R}_A$ is C, since A is the identity on $\mathcal{R}_A$. Also, since the range of C is contained in $\mathcal{R}_A$, we have $A \circ C = C$. Thus we obtain $\check{C} \circ \check{C} = C \circ A \circ C \circ A = C \circ C \circ A = C \circ A = \check{C}$, since $C \circ C = C$ anyway (it being an event in $\mathcal{L}_A$). Next we show symmetry for $\check{C}$. Note that for any m, $n \in \mathcal{R}$ we have $\langle m \mid m_{:C}\rangle\langle m_{:C} \mid n\rangle = \langle n \mid n_{:C}\rangle\langle n_{:C} \mid m\rangle$ by hypothesis. So, for any m, n, we obtain $\langle m_{:A} \mid m_{:A:C}\rangle\langle m_{:A:C} \mid n_{:A}\rangle = \langle n_{:A} \mid n_{:A:C}\rangle\langle n_{:A:C} \mid m_{:A}\rangle$. Multiply by $\langle m \mid m_{:A}\rangle$ and $\langle n_{:A} \mid n\rangle$ to obtain $\langle m \mid m_{:A}\rangle\langle m_{:A} \mid m_{:A:C}\rangle\langle m_{:A:C} \mid n_{:A}\rangle\langle n_{:A} \mid n\rangle$ for the left side; by symmetry the first two factors become $\langle m_{:A:C} \mid m_{:A:C:A}\rangle\langle m_{:A:C:A} \mid m\rangle = \langle m_{:A:C} \mid m\rangle$ (since $A \circ C = C$), and the other two become $\langle m_{:A:C} \mid m_{:A:C:A}\rangle\langle m_{:A:C:A} \mid n\rangle = \langle m_{:A:C} \mid n\rangle$. So, the left side becomes $\langle m \mid m_{:A:C}\rangle\langle m_{:A:C} \mid n\rangle$ and similarly the right becomes $\langle n \mid n_{:A:C}\rangle\langle n_{:A:C} \mid m\rangle$, i.e., we obtain symmetry for the map $\check{C}$. Now note that $A \circ \check{C} = A \circ C \circ A = C \circ A = \check{C}$ and so by the argument in Proposition 2, which used only the symmetry property, we have $\langle m \mid m_{:\check{C}}\rangle\langle m_{:\check{C}} \mid n\rangle = m(A)\langle m_{:A} \mid m_{:A:\check{C}}\rangle\langle m_{:A:\check{C}} \mid n\rangle$ for all n. So let $\langle m \mid m_{:\check{C}}\rangle = 0$. If $m_{:A:\check{C}} \neq \theta$, we obtain $m(A)\langle m_{:A} \mid m_{:A:\check{C}}\rangle = 0$; but $m_{:A:\check{C}} = m_{:A:C}$, so we have $m(A)\langle m_{:A} \mid m_{:A:C}\rangle = 0$. Since C is an event in $\mathcal{L}_A$ the second factor cannot be 0, and so $m_{:\check{A}} = \theta$, which implies $m_{:A:\check{C}} = \theta$—a contradiction. So $m_{:A:\check{C}}$ is θ after all, and so $m_{:A:C}$ is also; but this is just $m_{:\check{C}}$! Thus, the second condition in definition 1 is also valid and $\check{C}$ is an event. Since $A \circ \check{C} = \check{C}$, we have by the definition that $\check{C} \leq A$.

We shall elevate this situation to the state of an axiom in the following form:

(S) *Subspace principle*: Whatever axiom we impose on $\mathcal{L}$ shall also be assumed to hold for each $\mathcal{L}_A$.

It should be noted that the example at the end of Sec. 2 satisfies the subspace principle, but still has not enough structure; this appears to indicate that our principle is not really too strong.

In case we mention explicitly the "secondary" axioms derived from the subspace principle, we shall mark them with an asterisk, in order to remind the reader that we are not making an independent assumption.

So we have as a consequence:

Axiom 2*: For any $A \in \mathcal{L}$, and any maximal orthogonal set $\{m_i\}$ in $\mathcal{R}_A$ we have $\sum\langle m \mid m_i\rangle = 1$ for each $m \in \mathcal{K}_A$.

Remark: This, actually, could have been incorporated in our definition of an event. It is trivially verifiable for O and each A_m, while for I it reverts back to Axiom 2.

Proposition 5: Let $\{m_i\}$ be m.o. in $\mathcal{K}_A$ and $m \in \mathcal{P}$; then $m(A) = \sum\langle m \mid m_i\rangle$. In particular, $m \in \mathcal{K}_A$ iff $\sum\langle m \mid m_i\rangle = 1$.

Proof: For $n \in \mathcal{R}_A$ we have $\langle m \mid m_{:A}\rangle\langle m_{:A} \mid n\rangle = \langle m \mid n\rangle$; hence $m(A)\sum\langle m_{:A} \mid m_i\rangle = \sum\langle m \mid m_i\rangle$, and since $m_{:A} \in \mathcal{K}_A$ we have $\sum\langle m_{:A} \mid m_i\rangle = 1$.

4. COMPLEMENTS

We shall now impose further structure on $\mathcal{L}$. But first we need some preliminary analysis in which the results of the previous section play a useful role.

Proposition 6: For any $A \in \mathcal{L}$ we have $(\mathcal{K}_A^{\perp})^{\perp} = \mathcal{K}_A$. Therefore $\mathcal{R}_A = (c\mathcal{D}_A)^{\perp}$ also.

It is evident that $m \in \mathcal{R}_A$ implies $m \perp n$ for all $n \in \mathcal{K}_A^{\perp}$, i.e., $m \in (\mathcal{K}_A^{\perp})^{\perp}$. So we must prove the reverse; here we use Proposition 5. Take $\{m_i\}$ m.o. in $\mathcal{R}_A$ and enlarge by $\{n_j\}$ to obtain m.o. in $\mathcal{P}$. By Proposition 5 we have each $n_j \in \mathcal{K}_A^{\perp}$, because $\sum\langle m \mid m_i\rangle + \sum\langle m \mid n_j\rangle = 1$ for any m, while for $m \in \mathcal{R}_A$ the first term is 1; hence $m \perp n_j$, for all j. Now if $m \in (\mathcal{K}_A^{\perp})^{\perp}$, the second term is zero, hence $\sum\langle m \mid m_i\rangle = 1$ and again by Proposition 5 we get $m \in \mathcal{K}_A$.

We can now introduce complements of events; the basic property is, of course, that the complement of an event can occur with certainty in some state iff the event itself cannot occur and vice versa. Existence of complements must be postulated, as the example at the end of Sec. 2 shows.

Axiom 3: For each $A \in \mathcal{L}$ there exists an event A' such that $\mathcal{K}_{A'} = \mathcal{K}_A^{\perp}$.

Note that $\mathcal{K}_A^{\perp}$ is, by the above, the same as $c\mathcal{D}_A$ and so it follows that $\mathcal{D}_{A'} = c(\mathcal{K}_{A'}^{\perp}) = c\mathcal{K}_A$.

The proof of the next result is again simple and shall be omitted.

Theorem 4: Complementation has the following properties: $(A')' = A$, $A \leq B$ implies $B' \leq A'$, $O' = \mathrm{I}$, $\mathrm{I}' = O$, $A \vee A' = \mathrm{I}$, $A \wedge A' = 0$ (where $\vee$, $\wedge$ denote supremum and infimum respectively).

The concept of disjointness for two events A, B is introduced in the usual way as the condition that $A \leq B'$ (or equivalently $B \leq A'$). It follows immediately that it is equivalent to $\mathcal{R}_A \perp \mathcal{R}_B$, i.e., that $m_{:A} \perp n_{:B}$ for any m, $n \in \mathcal{P}$. We write $A \perp B$, as usual.

At this stage we must ask the question: since the elements of $\mathcal{P}$ are supposed to represent the (pure) states of the system, we should have an additive property, i.e., for any family $\{A_i\}$ of pairwise disjoint events for

which $\sup A_i = A$ exists, we have $mA = \sum mA_i$. This is indeed the case.

Theorem 5: Suppose that $A_i \perp A_j$ for $i \neq j$ and that $A = \sup\{A_i\}$ exists. Then, for each $m \in \mathcal{P}$ we have $mA = \sum mA_i$.

Proof: First note that if we adjoin A' to the given family we obtain one with pairwise disjoint members whose supremum is I. Next note that $mA + mA' = 1$ for any A. Because if we take $\{m_i\}$ and $\{n_j\}$ m. o. in $\mathcal{R}_A$, $\mathcal{R}_{A'}$ and assume $m \perp m_i$, n_j, we have $m \in \mathcal{R}_A^{\perp}$ (by Proposition 5). Hence $m \in \mathcal{R}_{A'}$ which means that $\{m, n_j\}$ is orthogonal, which is impossible; thus $\{m_i, n_j\}$ is m. o., and so $\sum \langle m \mid m_i \rangle + \sum \langle m \mid n_j \rangle = 1$, i. e., the desired result.

Thus, it suffices to obtain our result for the case where $A = \mathrm{I}$. Choose $\{m_{ij}\}_j$ m. o. in A_i and note that all the m_{ij} together form an orthogonal family. Now let $m \perp m_{ij}$ for all i, j; this implies $A_m \leq A_i'$ for all i, hence $A_i \leq A_m'$ and so $\mathrm{I} \leq A_m'$, but this is absurd. Thus $\{m_{ij}\}_{i,j}$ is m. o. in $\mathcal{P}$ and so $\sum_i \sum_j \langle m \mid m_{ij} \rangle = 1$, or $\sum_i mA_i = 1$.

We shall write, as usual, the supremum of disjoint events as a sum.

It is also proper at this point to ask whether the orthomodular law is valid: Given $A \leq B$, is $B = A + (B \wedge A')$? This is a simple consequence of Axiom 3*.

Theorem 6: Given $A \leq B$; then $B = A + (B \wedge A')$.

Proof: Consider $\mathcal{L}_B$; by Axiom 3* there exists an element of $\mathcal{L}_B$ (which must have the form $\hat{C}$) such that $\hat{A} + \hat{C} = \hat{B}$ ($= \mathrm{I}$ for $\mathcal{L}_B$!). We claim that $C = B \wedge A'$ and $A + C = B$. We have $\mathcal{R}_C = \mathcal{R}_{\hat{C}} = \mathcal{R}_{\hat{A}}^{\perp} = \{m \in \mathcal{R}_B \mid m \perp \mathcal{R}_{\hat{A}}\}$; since $\mathcal{R}_{\hat{A}} = \mathcal{R}_A$, we have $\mathcal{R}_C = \mathcal{R}_B \cap \mathcal{R}_A^{\perp} = \mathcal{R}_B \cap \mathcal{R}_{A'}$ which implies (by Proposition 1) that $C = B \wedge A'$. We note that $\hat{A} + \hat{C} = \hat{B}$ means that, if $\{m_i\}$ and $\{n_j\}$ are m. o. in $\mathcal{R}_{\hat{A}}$ and $\mathcal{R}_{\hat{C}}$ respectively, then $\{m_i, n_j\}$ is m. o. in the state space, i. e., $\mathcal{R}_B$. But since $\mathcal{R}_{\hat{A}} = \mathcal{R}_A$, $\mathcal{R}_{\hat{C}} = \mathcal{R}_C$, we see that $mA + mC = mB$ for all $m \in \mathcal{P}$; thus $m \perp \mathcal{R}_B$ iff $m \perp \mathcal{R}_A$ and $m \perp \mathcal{R}_C$, i. e., $\mathcal{R}_B^{\perp} = \mathcal{R}_A^{\perp} \cap \mathcal{R}_C^{\perp}$, or $\mathcal{R}_{B'} = \mathcal{R}_{A'} \cap \mathcal{R}_{C'}$. Therefore $B' = A' \wedge C'$; hence $B = A + C$.

5. COMPATIBILITY

There are several equivalent formulations of compatibility (or commutativity) for two events. We shall use the following:

Definition 4: The events A, B are compatible (or commute) iff there exist pairwise disjoint A_1, B_1, C such that $A = A_1 + C$, $B = B_1 + C$.

It follows that $C = A \wedge B$. It is worth noting in general, that if the infimum, say C, of A and B exists, then $\mathcal{R}_C = \mathcal{R}_A \cap \mathcal{R}_B$. Because evidently $\mathcal{R}_C \subseteq \mathcal{R}_A \cap \mathcal{R}_B$, while if $m \in \mathcal{R}_A \cap \mathcal{R}_B$ we have $A_m \leq A$, $A_m \leq B$ hence $A_m \leq C$, i. e., $m \in \mathcal{R}_C$ and $\mathcal{R}_A \cap \mathcal{R}_B \subseteq \mathcal{R}_C$.

Theorem 7: The events A, B are compatible iff for each m we have $m_{:A:B} = m_{:B:A}$; in such a case this state is also $m_{:A \wedge B}$.

Proof: First suppose A, B compatible and write $A = A_1 + C$, $B = B_1 + C$ with $C = A \wedge B$, $A_1 \perp B$, and $B_1 \perp A$. Since $A_1 \leq A$ we have by Proposition 2 that $m_{:A_1:A} = m_{:A:A_1}$, hence, also $m_{:B:A:A_1} = m_{:B:A_1:A}$. But $B \perp A_1$ and so $m_{:B} \perp \mathcal{R}_{A_1}$, i. e., $m_{:B:A_1} = \theta$; thus $m_{:B:A:A_1} = \theta$ also, or $m_{:B:A} \in \mathcal{R}_{A_1}^{\perp}$. If $m_{:B:A} \neq \theta$, then $m_{:B:A}(A) = 1$, so that $m_{:B:A}(A_1) + m_{:B:A}(C) = 1$, and since we just saw that the first is zero, we obtain $m_{:B:A}(C) = 1$. So we have $m_{:B:A} \in \mathcal{R}_C$, or $m_{:B:A:C} = m_{:B:A}$; but $C \leq A$ and so $m_{:B:A:C} = m_{:B:C}$, while $C \leq B$ also and so $m_{:B:C} = m_{:C}$. Thus we conclude that $m_{:B:A} = m_{:C}$. Now if $m_{:B:A} = \theta$, we can assume $m_{:B} \neq \theta$ [for otherwise $m(B) = 0$ and $C \leq B$ implies $m(C) = 0$, i. e., $m_{:C} = \theta$ and again we get $m_{:B:A} = m_{:C}$]. So $m_{:B} \in \mathcal{R}_A$ and $m_{:B}(C) = 0$ or $m_{:B:C} = \theta$; again $C \leq B$ implies $m_{:C} = \theta$ and we end up with $m_{:B:A} = m_{:C}$. So in all cases we have $m_{:B:A} = m_{:A \wedge B}$; as the role of A, B is symmetric, we also have $m_{:A:B} = m_{:A \wedge B}$.

Now for the converse: Suppose $m_{:A:B} = m_{:B:A}$ for all m; we shall first establish the existence of $A \wedge B$, which has to be the map $m \to m_{:A:B}$ by the previous argument. Write C for the map $A \circ B$ ($= B \circ A$) and note that idempotency is trivial: $C \circ C = A \circ B \circ A \circ B = A \circ A \circ B \circ B = A \circ B = C$. Symmetry takes a little longer. Note that $\langle m \mid m_{:A} \rangle \langle m_{:A} \mid m_{A:B} \rangle = \langle m_{:A:B} \mid m_{:A:B:A} \rangle \langle m_{A:B:A} \mid m \rangle = \langle m_{:A:B} \mid m_{:A:A:B} \rangle \langle m_{A:A:B} \mid m \rangle = \langle m_{:A:B} \mid m_{:A:B} \rangle \langle m_{:A:B} \mid m \rangle = \langle m \mid m_{:A:B} \rangle$, and so we have $\langle m \mid m_{:A:B} \rangle \langle m_{:A:B} \mid n \rangle = \langle m \mid m_{:A} \rangle \langle m_{:A} \mid m_{:A:B} \rangle \langle m_{:A:B} \mid n \rangle = \langle m \mid m_{:A} \rangle \langle n \mid n_{:B} \rangle \times \langle n_{:B} \mid m_{:A} \rangle = \langle n \mid n_{:B} \rangle \langle n_{:B} \mid n_{:B:A} \rangle \langle n_{:B:A} \mid m \rangle = \langle n \mid n_{:B:A} \rangle \times \langle n_{:B:A} \mid m \rangle = \langle n \mid n_{:A:B} \rangle \langle n_{:A:B} \mid m \rangle$ which is precisely what we want. Finally, if $\langle m \mid m_{:A:B} \rangle = 0$, we have by our very first remark that either $\langle m \mid m_{:A} \rangle = 0$ or $\langle m_{:A} \mid m_{:A:B} \rangle = 0$, i. e., either $m = \theta$, or $m_{:A} = \theta$ or $m_{:A:B} = \theta$ and so the composite map $A \circ B$ is indeed an event.

By the definition of $\leq$ we have at once $C \leq A$, $C \leq B$, i. e., $\mathcal{R}_C \subseteq \mathcal{R}_A \cap \mathcal{R}_B$. On the other hand, $m \in \mathcal{R}_A \cap \mathcal{R}_B$ implies $m = m_{:A}$ and $m = m_{:B}$, hence $m_{:A:B} = m_{:B} = m$, i. e., $m \in \mathcal{R}_C$ and this shows $\mathcal{R}_C = \mathcal{R}_A \cap \mathcal{R}_B$, so that indeed $C = A \wedge B$. To show that A, B are compatible we must show that $A \wedge (A \wedge B)' \perp B$, $B \wedge (A \wedge B)' \perp A$. We verify the first, since the role of A, B is symmetric. Let $m[A \wedge (A \wedge B)'] = 1$, which implies $mA = 1$, and $m(A \wedge B) = 0$. Then $m_{:A} = m$, and $m_{:A:B} = m_{:B:A}$, gives $m_{:B} = m_{:B:A}$. Let $m_{:B} \neq \theta$; then $m_{:B}(A) = m_{:B:A}(A) = 1$, and so $m_{:B} \in \mathcal{R}_A$. Thus $m_{:B} \in \mathcal{R}_{A \wedge B}$, hence $m_{:B}(A \wedge B) = 1$. But since $A \wedge B \leq B$ we have $m(A \wedge B) = m(B) m_{:B}(A \wedge B)$, or $m(A \wedge B) = m(B)$, which implies $m(B) = 0$; but this is impossible since we assumed $m_{:B} \neq \theta$. This means that $m_{:B} = \theta$, or indeed $m(B) = 0$, i. e., that $A \wedge (A \wedge B)' \leq B'$. QED

6. THE STATES

In general, a state of a logic $\mathcal{L}$ is defined as a map $m: \mathcal{L} \to [0, 1]$ such that $m(0) = 0$, $m(\mathrm{I}) = 1$, and $m(A) = \sum m(A_i)$ for A_i pairwise disjoint with A their supremum. In the space of all maps $\mathcal{L} \to \mathcal{R}$, the states evidently form a convex set whose extreme points are, by definition, the pure states. Implicit in our analysis is the hypothesis that these extreme points are in a one-to-one correspondence with the elements of $\mathcal{P}$. Naturally this requirement is expected to limit severely the range of candidates for the functions $\langle \mid \rangle$ which determine the structure of $\mathcal{L}$.

Proposition 7: If $\{m_i\}$ is m. o. in $\mathcal{R}_A$ then $A = \sum A_i$,

where $A_i = A_{m_i}$. In particular, if f is any state of $\mathcal{L}$, then $f(A) = \sum f(A_i)$.

Proof: Clearly $A_i \leq A$ for all i. Now, if $A_i \leq B$ we have $m_i \in \mathcal{R}_B$, and so there exist $n_j \in \mathcal{P}$ with $\{m_i, n_j\}$ m. o. in $\mathcal{R}_B$. It follows that $\sum \langle m | m_i \rangle + \sum \langle m | n_j \rangle = m(B)$ for all $m \in \mathcal{P}$; but $m \in \mathcal{R}_A$ implies the first term is 1, hence $m(B) = 1$ and $m \in \mathcal{R}_B$, i. e., $\mathcal{R}_A \subseteq \mathcal{R}_B$. Thus, $A_i \leq B$ implies $A \leq B$ and so $A = \sum A_i$. The rest follows from the definition of states.

This means that a state f is completely determined by its values on the atoms of $\mathcal{L}$, and since these are in a 1 : 1 correspondence with $\mathcal{P}$, we may consider f as a function on $\mathcal{P}$. Evidently $\sum_{m \in Q} f(m) = 1$ for each m. o. set $Q \subseteq \mathcal{P}$. The converse of this also holds.

Proposition 8: Let f be a map $\mathcal{P} \to [0, 1]$ such that for each m. o. set $Q \subseteq \mathcal{P}$ we have $\sum_{m \in Q} f(m) = 1$. Then, given $A \in \mathcal{L}$ and $\{m_i\}$ m. o. in $\mathcal{R}_A$, the number $\sum_i f(m_i)$ is independent of the particular m. o. set employed; if we write it as $\hat{f}(A)$, the map $\hat{f} : \mathcal{L} \to [0, 1]$ obtained is a state of $\mathcal{L}$ such that $\hat{f}(A_m) = f(m)$.

Proof: The last part follows easily, provided $\hat{f}(A)$ is well defined. Because, as we saw in the proof of Theorem 5, given $A = \sum A_i$ and $\{m_{ij}\}_j$ m. o. in $\mathcal{R}_{A_i}$, then $\{m_{ij}\}_{i,j}$ is m. o. in $\mathcal{R}_A$ and $\hat{f}(A) = \sum_{i,j} f(m_{ij}) = \sum_i \sum_j f(m_{ij}) = \sum_i \hat{f}(A_i)$, i. e., $\hat{f}$ is a state. So take two m. o. sets $\{m_i\}$, $\{n_j\}$ in $\mathcal{R}_A$ and some m. o. set $\{p_k\}$ in $\mathcal{R}_{A'} = \mathcal{R}_A^{\perp}$. Then $\{m_i, p_k\}$ and $\{n_j, p_k\}$ are m. o. in $\mathcal{P}$, hence we have $\sum f(m_i) + \sum f(p_k) = 1 = \sum f(n_j) + \sum f(p_k)$, i. e., $\sum f(m_i) = \sum f(n_j)$, the desired independence.

Let $\mathcal{H}$ be the (convex) set of all maps f: $\mathcal{P} \to [0, 1]$ such that $\sum_{m \in Q} f(m) = 1$ for each m. o. set Q in $\mathcal{P}$, and $\mathcal{H}_e$ the set of extreme points of $\mathcal{H}$. The implicit restriction mentioned previously becomes:

(P) There exists a 1 : 1 correspondence $\mathcal{P} \ni m \leftrightarrow f_m \in \mathcal{H}_e$ such that $f_m(N) = \langle m | n \rangle$.

This, even though quite clearly formulated, does not appear simple to verify or contradict; the calculation of extreme points is quite complicated. We shall now see how this goes through for the systems mentioned in Sec. 1, and then we shall exploit this analysis to show that no system of states for which (*P*) holds can be finite.

Example 1: (revisited) Recall that we have $\langle m | n \rangle = \delta_{mn}$, and we have verified that the events are in a 1 : 1 correspondence with the subsets of $\mathcal{P}$. It is clear that this correspondence is the one we developed in Sec. 2 for any $\mathcal{L}$, namely $A \leftrightarrow \mathcal{R}_A$. Axiom 3 is easily seen to hold, with A' just the set complement of A, and the subspace principle also holds. Our $\mathcal{L}$ is just the atomic (complete) Boolean algebra of all subsets of $\mathcal{P}$, and thus its pure states are just the Dirac measures on $\mathcal{P}$. So, we see that the required 1 : 1 correspondence between $\mathcal{P}$ and $\mathcal{H}_e$ is there.

Example 2: (revisited) We have already seen that any projection gives rise to an event in our $\mathcal{L}$. Now we verify the converse. Take any $A \in \mathcal{L}$ and $\{m_{\psi_i}\}$ a m. o. set in $\mathcal{R}_A$; thus $\{\psi_i\}$ is orthonormal in H. We have $m_\psi \in \mathcal{R}_A$ iff $\sum \langle m_\psi | m_{\psi_i} \rangle = 1$, i. e., iff $\sum |(\psi | \psi_i)|^2 = 1$, which means $\psi = \sum (\psi | \psi_i) \psi_i$. Write M_A for the (closed) subspace spanned by the $\{\psi_i\}$ and P_A for the corresponding projection. For any $m_\psi \in \mathcal{P}$ we have $m_\psi \in \mathcal{D}_A$ iff ψ is not orthogonal to $\mathcal{R}_A$, i. e., iff $\psi \notin M_A^{\perp}$; further, we know that $m_{\psi:A}$ maximizes $\langle m_\psi | n \rangle$ as n varies in $\mathcal{R}_A$. Thus we seek for the maximum of $\langle m_\psi | m_x \rangle$ as x varies in M_A, which is obtained for $x = P_A\psi / ||P_A\psi||$. This shows that the map A precisely corresponds to P_A as required. It is also clear that A' corresponds to the projection on $M_A^{\perp}$, and that the subspace principle is valid.

Invoking Gleason's theorem we see that the pure states are indeed in a 1 : 1 correspondence with the elements of $\mathcal{P}$ as demanded in requirement (*P*) above.

For the most part, the analysis that follows is valid in any logic; it is only at the end that (*P*) is invoked. Let us assume that a finite logic exists, which is not Boolean. Then there must exist one, say $\mathcal{L}$, with the least number of elements in $\mathcal{P}$. But then, by the subspace principle, each $\mathcal{L}_A$ (for $A \neq \mathrm{I}$) is Boolean.

Lemma 2: Suppose that for $A \neq \mathrm{I}$ each $\mathcal{L}_A$ is Boolean. Then, if $\mathcal{B}$ is a maximal Boolean algebra in $\mathcal{L}$, no element $A \notin \mathcal{B}$ can commute even with a single element $B \in \mathcal{B}$ ($B \neq 0$ or I).

Proof: First we consider the simple case of the relation $A \leq B$. If this happens, then A commutes with all $C \leq B$ because $\mathcal{L}_B$ is Boolean. Now take $B_1 \in \mathcal{B}$ and write $B_1 = (B \wedge B_1) + (B' \wedge B_1)$; we have $A \perp B' \wedge B_1$ and A commuting with $B \wedge B_1$. It follows that A commutes with their sum B_1. For convenience, introduce a short notation: let $X \perp Y$ with $A \perp Y$ and A compatible with X. We have $A = A_1 + C$, $X = X_1 + C$, and $A_1 \perp X_1$. So $X + Y = X_1 + Y + C$ and all we need is $C \perp X_1 + Y$, $A_1 \perp X_1 + Y$. But $C \leq X$, hence $C \perp Y$ and $C \perp X_1$ by hypothesis, so $C \perp X_1 + Y$; also $A_1 \perp X_1$ by hypothesis and $A_1 \leq A \perp Y$, so $A_1 \perp Y$ too, hence $A_1 \perp X_1 + Y$. So we see that if $A \leq B$ for some $B \in \mathcal{B}$, then A is compatible with all $B_1 \in \mathcal{B}$; since $\mathcal{B}$ is maximal such an A cannot be outside $\mathcal{B}$.

Now for the general case. Suppose A is compatible with some $B \in \mathcal{B}$; by our argument above we have $A \wedge B \in \mathcal{B}$, hence $B \wedge (A \wedge B)' \in \mathcal{B}$ also. But $A \wedge (A \wedge B)'$ is disjoint from $B \wedge (A \wedge B)'$ by hypothesis, hence $\leq$ to its complement which is in $\mathcal{B}$. Thus $A \wedge (A \wedge B)' \in \mathcal{B}$ and finally $A \in \mathcal{B}$ being the sum of two events in $\mathcal{B}$.

This lemma shows that in a logic $\mathcal{L}$ for which all $\mathcal{L}_A$ ($A \neq \mathrm{I}$) are Boolean, the maximal Boolean algebras have no elements in common except O, I; also that events in distinct maximal Boolean algebras are not related by $\leq$, $\perp$.

Thus, such an $\mathcal{L}$ is a disjoint union of its maximal Boolean subalgebras. We apply this to the logic mentioned previously, the one for which $\mathcal{P}$ has the least number of elements. We then have $\mathcal{L} = \cup_{i=1}^{k} \mathcal{B}_i$, and we observe that each state of $\mathcal{L}$ is determined by a family $(m_i)_{i=1,\cdots,k}$ of states, one for each $\mathcal{B}_i$. Each m_i can be chosen independently because there is no relation between the elements of distinct $\mathcal{B}_i$; the same goes for pure states. Note that since the $\mathcal{B}_i$ are Boolean, the values of the pure states on the atoms are either 0 or 1. But by virtue of requirement (*P*) this means that each $\langle m | n \rangle$ (for $m, n \in \mathcal{P}$) is either 0 or 1. Since $\langle m | n \rangle \neq 1$ for $m \neq n$, we have all "off diagonal" probabilities

zero, and $\mathcal{L}$ is, after all, Boolean—contrary to our hypothesis.

So we have completed the proof of:

Theorem 8: There is no finite logic derivable from a state space by means of Axioms 1—3, and the subspace principle ($\mathcal{S}$) satisfying requirement (P) on the behavior of pure states, unless it is Boolean.

We have as yet been unable to determine whether a non-Boolean countable model exists.

7. ORTHOCOMPLETENESS

So far we have made no assumptions with regard to the existence of suprema or infima—even for disjoint events. It is not hard to see that if $A=\inf\{A_i\}$ exists then $\mathcal{R}_A=\cap\mathcal{R}_{A_i}$ (we have actually used part of this in Sec. 5). Evidently $\mathcal{R}_A\subseteq\cap\mathcal{R}_{A_i}$, so let $m\in\cap\mathcal{R}_{A_i}$; then $A_m\leq A_i$, which implies $A_m\leq A$, or $mA=1$ and $m\in\mathcal{R}_A$. The meaning of this is, of course, that the conjunction of a family of events occurs with certainty in a state iff each event occurs with certainty in that state.

It is perhaps interesting to see that the hypothesis of orthocompleteness (i.e., that every family of pairwise disjoint events admits a supremum) can be used to replace several of our assumptions.

First note that it is sufficient to assume orthocompleteness for atoms only.

Proposition 9: If for any orthogonal family $\{m_i\}$, the supremum $\sum A_{m_i}$ exists, then $\mathcal{L}$ is orthocomplete.

Proof: Recall that if $\{n_j\}$ is a m.o. set in $\mathcal{R}_A$, then $A=\sum A_{n_j}$ (Proposition 7). Now consider any disjoint family $\{A_i\}$ with $\{m_{ij}\}_j$ m.o. in $\mathcal{R}_{A_i}$. Let $A=\sum_{i,j}A_{m_{ij}}$, so that $A\geq A_i$ for all i. We then see, following the argument in Proposition 7, that if $B\geq A_i$ for all i, then $m_{ij}\in\mathcal{R}_B$ and so $A\leq B$.

In view of what we have already developed, orthocompleteness is equivalent to:

Given any orthogonal set $\{m_i\}$, there exists an event A such that $\mathcal{R}_A=\{m\,|\,\sum\langle m\,|\,m_i\rangle=1\}$. This is roughly the converse of Proposition 5, and provides a characterization of the various possible $\mathcal{R}_A$.

It is this characterization which can replace Axioms 2, 3 and the subspace principle. Technically it appears as an improvement, but it lacks an immediate physical interpretation. In exact terms it can be stated as follows:

(0) Given any m.o. $\{m_i\}$ in $\mathcal{R}_A$, we have $\mathcal{R}_A=\{m\,|\,\sum\langle m\,|\,m_i\rangle=1\}$ and vice versa, given any orthogonal $\{n_j\}$, there exists an event B such that $\{m\,|\,\sum\langle m\,|\,n_j\rangle=1\}=\mathcal{R}_B$.

Let us recall that the partial ordering of $\mathcal{L}$ and its properties did not require the use of Axioms 2, 3 or (S). We now use (0) to obtain the existence of complements.

First note that Proposition 5 is incorporated in (0) and so Proposition 6 is still valid: For any event A we have $(\mathcal{R}_A^{\perp})^{\perp}=\mathcal{R}_A$. Now take any m.o. set $\{m_i\}$ in $\mathcal{R}_A$; we claim that the event A' associated to it is such that $\mathcal{R}_{A'}=\mathcal{R}_A^{\perp}$. We must show that $\sum\langle m\,|\,m_i\rangle=1$ iff $m\in\mathcal{R}_A$. Choose any m.o. set $\{n_j\}$ in $\mathcal{R}_A$ and note that $\{m_i,n_j\}$ is m.o. in $\mathcal{P}$; because if $\langle m\,|\,n_j\rangle=0$ for all j, then $m\in\mathcal{R}_A^{\perp}$ (by Prop. 5) hence cannot be $\perp$ all m_i. Thus, $\sum\langle m\,|\,m_i\rangle+\sum\langle m\,|\,n_j\rangle=1$ for all m [by (0)]; since $m\in\mathcal{R}_A$ iff $\langle m\,|\,n_j\rangle=0$ for all j we are done.

Now we verify orthomodularity. Take $A\leq B$, any m.o. set $\{m_i\}$ in $\mathcal{R}_A$ and enlarge it to a m.o. set $\{m_i,n_j\}$ for $\mathcal{R}_B$. Write C for the event corresponding to $\{n_j\}$. Evidently $A\perp C$, and all we have to do is to show $A+C=B$. Now we know by our remarks at the beginning of the section that $A+C$ exists, and it is of course $\leq B$. By the argument in Theorem 5 we also know that $\{m_i,n_j\}$ is a m.o. set in $\mathcal{R}_{A+C}$; but then $A+C=B$ and we are done.

[1]B. Mielnik, Comm. Math. Phys. **9**, 55 (1968).
[2]B. Mielnik, Comm. Math. Phys. **15**, 1 (1969).
[3]E. B. Davies, Comm. Math. Phys. **15**, 227 (1969).
[4]C. M. Edwards, Comm. Math. Phys. **16**, 207 (1970).
[5]E. B. Davies and J. T. Lewis, Comm. Math. Phys. **17**, 239 (1970).
[6]S. P. Gudder, Comm. Math. Phys. **29**, 249 (1973).
[7]B. Mielnik, Comm. Math. Phys. **37**, 221 (1974).

Algebraically special $\mathcal{H}$-spaces*

C. W. Fette

The Pennsylvania State University at McKeesport, McKeesport, Pennsylvania 15132

Allen I. Janis† and Ezra T. Newman†

University of Pittsburgh, Pittsburgh, Pennsylvania 15260
(Received 22 September 1975)

All diverging algebraically special solutions of the complex vacuum Einstein equations which are left (or right) conformally flat ($\mathcal{H}$-spaces) are found explicitly. These metrics contain four arbitrary functions of two variables.

I. INTRODUCTION

Recently a four-complex-dimensional manifold known as $\mathcal{H}$-space has been introduced into general relativity.[1-5] It arose first in the study of asymptotically flat solutions of the Einstein or Maxwell—Einstein equations as the manifold of asymptotically shear-free null cones of the analytically extended asymptotically flat (physical) space—time. The study of $\mathcal{H}$-space has already proved to be a powerful tool for the analysis of the real space—time with which an $\mathcal{H}$-space is associated. A complex center-of-mass world line[5] can be defined in $\mathcal{H}$-space, which leads to a definition of center-of-mass motion in general relativity. An intrinsic angular momentum can also be associated with this center-of-mass world line. A magnetic moment similarly arises from a complex center-of-charge world line in $\mathcal{H}$-space. It has been shown for stationary space—times that if these two world lines coincide, the Dirac value ($g=2$) for the gyromagnetic ratio immediately follows. It appears that an entire theory of equations of motion follows from these considerations. More recently, Penrose has argued from another point of view that $\mathcal{H}$-space should be viewed as a nonlinear graviton.

One can show that $\mathcal{H}$-space has the following properties.[3,4] It is a four-complex-dimensional manifold with a nondegenerate *complex* "Riemannian" metric on it. In addition to satisfying the (complex) vacuum Einstein equations the metric is such that the self-dual (or anti-self-dual) part of the Weyl tensor is zero. [In terms of a spinor description this means that $\Psi_{ABCD}=0$ (or $\tilde{\Psi}_{A'B'C'D'}=0$). Note that due to the complex nature of the space, $\tilde{\Psi}$ is not the complex conjugate of Ψ nor is the anti-self-dual part of the Weyl tensor the conjugate of the self-dual part; they are independent of each other.] An $\mathcal{H}$-space with $\Psi_{ABCD}=0$ is referred to as left flat while when $\tilde{\Psi}_{A'B'C'D'}=0$ it is right flat.

It is the purpose of this paper to find explicit examples of $\mathcal{H}$-spaces. In particular we obtain all algebraically special $\mathcal{H}$-spaces with nonvanishing divergence. In Sec. II the spin-coefficient form of the vacuum Einstein equations is generalized to include complex manifolds, and hence $\mathcal{H}$-space, and then integrated under the conditions of algebraic specialness. The further specialization to types III, N, and D is given in Sec. III. In Sec. IV we conclude with a discussion of some unsolved special problems related to $\mathcal{H}$-space.

II. ALGEBRAICALLY SPECIAL $\mathcal{H}$-SPACES

The spinor (or spin-coefficient) formalism[6] provides a convenient framework for generalizing the Einstein equations to complex manifolds. At each point of the four-complex-dimensional manifold we introduce two spin spaces, S and $\tilde{S}$, which are independent of each other, and a normalized basis in each space (π^A,λ^A), $(\tilde{\pi}^{A'},\tilde{\lambda}^{A'})$ so that $\pi_A\lambda^A=\tilde{\pi}_{A'}\tilde{\lambda}^{A'}=0$. By identifying the tangent space at a point in $\mathcal{H}$ with $S\otimes\tilde{S}$ at that point we obtain the complex null tetrad

$$l^\mu=\sigma^\mu_{AA'}\pi^A\tilde{\pi}^{A'},\quad n^\mu=\sigma^\mu_{AA'}\lambda^A\tilde{\lambda}^{A'},\quad m^\mu=\sigma^\mu_{AA'}\pi^A\tilde{\lambda}^{A'},$$
$$\tilde{m}^\mu=\sigma^\mu_{AA'}\lambda^A\tilde{\pi}^{A'},$$

the σ's being the Infeld—van der Waerden symbols. (We emphasize that, in the present formalism, a tilde does not indicate complex conjugation. The usual spin-coefficient formalism would result from specializing the present work to the case where all quantities with tildes are the complex conjugates of the corresponding quantities without tildes.)

The spin coefficients and intrinsic derivatives used in the spin-coefficient formalism are defined in the usual way, but again quantities that in the usual formalism are complex conjugates of each other are now independent quantities. (For example the operators $\delta\equiv m^\mu\nabla_\mu$ and $\tilde{\delta}\equiv\tilde{m}^\mu\nabla_\mu$ as well as the coefficients $\rho\equiv l_{\mu;\nu}m^\mu\tilde{m}^\nu$ and $\tilde{\rho}\equiv l_{\mu;\nu}\tilde{m}^\mu m^\nu$ are independent quantities. The gradient operator, of course, refers to the complex coordinates of $\mathcal{H}$-space.)

The complete set of complexified spin-coefficient equations now consists of the usual spin-coefficient Bianchi identities) with all complex-conjugated quantities replaced by tilded quantities, plus the equations obtained by interchanging the corresponding tilded and untilded quantities. [For example, the equation $D\Psi_1-\bar{\delta}\Psi_0=-3\kappa\Psi_2+(2\epsilon+4\rho)\Psi_1+(\pi-4\alpha)\Psi_0$ must be replaced by the two equations

$$D\Psi_1-\tilde{\delta}\Psi_0=-3\kappa\Psi_2+(2\epsilon+4\rho)\Psi_1+(\pi-4\alpha)\Psi_0$$

and

$$D\tilde{\Psi}_1-\delta\tilde{\Psi}_0=-3\tilde{\kappa}\tilde{\Psi}_2+(2\tilde{\epsilon}+4\tilde{\rho})\tilde{\Psi}_1+(\tilde{\pi}-4\tilde{\alpha})\tilde{\Psi}_0.\ ]$$

By setting the Φ_{mn} and Λ equal to zero (equivalent to $R_{\mu\nu}=0$) in the complexified spin-coefficient equations,

we have a set of first-order equations equivalent to the complex vacuum Einstein equations. In order to further restrict the complex space to $\mathcal{H}$-space, it is necessary to set

$$\Psi_0=\Psi_1=\Psi_2=\Psi_3=\Psi_4=0 \tag{2.1}$$

for left-flat spaces, or

$$\tilde{\Psi}_0=\tilde{\Psi}_1=\tilde{\Psi}_2=\tilde{\Psi}_3=\tilde{\Psi}_4=0$$

for right-flat spaces. This follows from the fact that

$$\Psi_0=-C^-_{\alpha\beta\gamma\delta}l^\alpha m^\beta l^\gamma m^\delta, \quad \tilde{\Psi}_0=-C^+_{\alpha\beta\gamma\delta}l^\alpha \tilde{m}^\beta l^\gamma \tilde{m}^\delta,$$
$$\Psi_1=-C^-_{\alpha\beta\gamma\delta}l^\alpha n^\beta l^\gamma m^\delta, \quad \tilde{\Psi}_1=-C^+_{\alpha\beta\gamma\delta}l^\alpha n^\beta l^\gamma \tilde{m}^\delta,$$
$$\Psi_2=-C^-_{\alpha\beta\gamma\delta}\tilde{m}^\alpha n^\beta l^\gamma m^\delta, \quad \tilde{\Psi}_2=-C^+_{\alpha\beta\gamma\delta}m^\alpha n^\beta l^\gamma \tilde{m}^\delta,$$
$$\Psi_3=-C^-_{\alpha\beta\gamma\delta}\tilde{m}^\alpha n^\beta l^\gamma n^\delta, \quad \tilde{\Psi}_3=-C^+_{\alpha\beta\gamma\delta}m^\alpha n^\beta l^\gamma n^\delta,$$
$$\Psi_4=-C^-_{\alpha\beta\gamma\delta}\tilde{m}^\alpha n^\beta \tilde{m}^\gamma n^\delta, \quad \tilde{\Psi}_4=-C^+_{\alpha\beta\gamma\delta}m^\alpha n^\beta m^\gamma n^\delta,$$

with $C^-_{\alpha\beta\gamma\delta}$ and $C^+_{\alpha\beta\gamma\delta}$ being the self-dual and anti-self-dual parts of the Weyl tensor, respectively:

$$C^-_{\alpha\beta\gamma\delta}=\tfrac{1}{2}(C_{\alpha\beta\gamma\delta}-iC^*_{\alpha\beta\gamma\delta}), \quad C^+_{\alpha\beta\gamma\delta}=\tfrac{1}{2}(C_{\alpha\beta\gamma\delta}+iC^*_{\alpha\beta\gamma\delta}).$$

Simply for definiteness we chose to work with left-flat spaces rather than the right-flat ones, so that we now impose Eqs. (2.1) on the spin-coefficient form of the complexified Einstein equations.

The definition of left-flat $\mathcal{H}$-space implies, for the spinor components of the Weyl tensor, that $\Psi_{ABCD}=0$, which further implies that the unprimed spin space is parallely propagated. Though it is possible to choose the basis spinors π^A and λ^A so that they are parallely propagated, it is much more convenient for us here not to do so. The reason will be apparent later.

The definition of an algebraically special $\mathcal{H}$-space is a simple modification of the usual definition. Since $\tilde{\Psi}_{A'B'C'D'}$ can always be written in the form

$$\tilde{\Psi}_{A'B'C'D'}=\tilde{\alpha}_{(A'}\tilde{\beta}_{B'}\tilde{\gamma}_{C'}\tilde{\delta}_{D')},$$

we define algebraic specialness by the equality of two of the principal spinors, i.e., by

$$\Psi_{A'B'C'D'}=\tilde{\alpha}_{(A'}\tilde{\alpha}_{B'}\tilde{\gamma}_{C'}\tilde{\delta}_{D')}. \tag{2.2}$$

If the basis spinor $\tilde{\pi}_{A'}$ is chosen to be the repeated principal spinor, this leads immediately to the conditions that

$$\tilde{\Psi}_0=\tilde{\Psi}_1=0. \tag{2.3}$$

We hereafter adopt these conditions.

Next we wish to choose the spinor basis (and thus the tetrad) and a coordinate system so that we simplify the spin-coefficient equations. The allowed tetrad transformations are induced by transformations of the spinor basis (at each point of $\mathcal{H}$) of the form

$$\pi^{*A}=a\pi^A+b\lambda^A, \quad \tilde{\pi}^{*A'}=\tilde{a}\tilde{\pi}^{A'}+\tilde{b}\tilde{\lambda}^{A'},$$
$$\lambda^{*A}=c\pi^A+d\lambda^A, \quad \tilde{\lambda}^{*A'}=\tilde{c}\tilde{\pi}^{A'}+\tilde{d}\tilde{\lambda}^{A'}, \tag{2.4}$$

with

$$\begin{vmatrix} a & b \\ c & d \end{vmatrix}=\begin{vmatrix} \tilde{a} & \tilde{b} \\ \tilde{c} & \tilde{d} \end{vmatrix}=1.$$

In order to preserve Eqs. (2.3), we have immediately that

$$\tilde{b}=0. \tag{2.5}$$

We find (after a great deal of effort) that, by writing out the appropriate transformation equations for the spin coefficients and examining their integrability conditions, we may choose the tetrad such that the following spin-coefficient relations are satisfied[7]:

$$\kappa=\epsilon=\pi=\tau=\lambda=0,$$
$$\tilde{\kappa}=\tilde{\epsilon}=\tilde{\pi}=\tilde{\tau}=\tilde{\sigma}=0, \tag{2.6}$$
$$\rho-\tilde{\rho}=\alpha+\tilde{\beta}=\tilde{\alpha}+\beta=0.$$

Note that had we chosen the unprimed spinor basis to be parallely propagated this simplification could not have been achieved, and most importantly l_μ would not have been geodesic and a gradient.

The transformations (2.4), when (2.3) and (2.6) are imposed, are restricted by the following conditions:

$$\tilde{b}=c=d-a^{-1}=\tilde{d}-\tilde{a}^{-1}=0,$$
$$D(b/a)-\rho(b/a)=0, \quad D\ln a-(\tilde{\beta}-\tilde{\delta}\ln a)(b/a)=0,$$
$$\tilde{\delta}(b/a)+2\tilde{\beta}(b/a)=0, \quad D\ln\tilde{a}+(\tilde{\beta}+\tilde{\delta}\ln\tilde{a})(b/a)=0,$$
$$\tilde{\delta}\ln(a\tilde{a})=0, \quad \Delta(b/a)-2\gamma(b/a)-\nu(b/a)^2-\rho\tilde{a}\tilde{c}=0,$$
$$\delta\ln(a\tilde{a})+\Delta(b/a)+[\Delta\ln(a\tilde{a})+\tilde{\gamma}-\gamma+\mu](b/a)=0,$$
$$D(\tilde{a}\tilde{c})+[\tilde{\delta}(\tilde{a}\tilde{c})+2\tilde{\beta}\tilde{a}\tilde{c}+\tilde{\mu}](b/a)=0.$$

Since l_μ is a gradient and tangent to a null geodesic, one can choose the scalar function of which it is the gradient as one of the coordinates, u, and the affine length along l^μ as another coordinate r. Two further coordinates, ζ and $\tilde{\zeta}$ (constant along each geodesic), label the geodesics. This leads to the form for the tetrad:

$$l_\mu=u_{,\mu}=\delta^0_\mu, \quad l^\mu=\frac{dz^\mu}{dr}=\delta^\mu_1, \tag{2.7a}$$

$$n^\mu=\delta^\mu_0+U\delta^\mu_1+X^k\delta^\mu_k, \tag{2.7b}$$

$$m^\mu=\omega\delta^\mu_1+\xi^k\delta^\mu_k, \tag{2.7c}$$

$$\tilde{m}^\mu=\tilde{\omega}\delta^\mu_1+\tilde{\xi}^k\delta^\mu_k, \tag{2.7c̃}$$

($k=2, 3$) and hence

$$D=\frac{\partial}{\partial r},$$
$$\Delta=\frac{\partial}{\partial u}+U\frac{\partial}{\partial r}+X^k\frac{\partial}{\partial z^k},$$
$$\delta=\omega\frac{\partial}{\partial r}+\xi^k\frac{\partial}{\partial z^k}, \tag{2.8}$$
$$\tilde{\delta}=\tilde{\omega}\frac{\partial}{\partial r}+\tilde{\xi}^k\frac{\partial}{\partial z^k}.$$

The full set of spin-coefficient equations for algebraically special $\mathcal{H}$-space can now be written as follows:

Field equations

$$D\rho=\rho^2, \tag{2.9a}$$

$$D\sigma=2\rho\sigma, \tag{2.9b}$$

$$D\beta=\rho\beta-\sigma\tilde{\beta}, \tag{2.9c}$$

$$D\tilde{\beta}=\rho\tilde{\beta}, \tag{2.9c̃}$$

$$D\gamma=0, \tag{2.9d}$$

$$D\tilde{\gamma}=\tilde{\Psi}_2, \tag{2.9d̃}$$

$$D\tilde{\lambda}=\rho\tilde{\lambda}+\sigma\tilde{\mu}, \tag{2.9ẽ}$$

$$D\mu=\rho\mu, \tag{2.9f}$$

$$D\tilde{\mu}=\rho\tilde{\mu}+\tilde{\Psi}_2, \tag{2.9f̃}$$

$$D\nu=0, \tag{2.9g}$$

$$D\tilde{\nu}=\tilde{\Psi}_3, \tag{2.9g̃}$$

$$\tilde{\delta}\nu=2\tilde{\beta}\nu, \tag{2.9h}$$

$$\Delta\tilde{\lambda}-\delta\tilde{\nu}=(\gamma-\mu-3\tilde{\gamma}-\tilde{\mu})\tilde{\lambda}-2\beta\tilde{\nu}-\tilde{\Psi}_4, \tag{2.9h̃}$$

$$\delta\rho-\tilde{\delta}\sigma=4\tilde{\beta}\sigma, \tag{2.9i}$$

$$\tilde{\delta}\rho=0, \tag{2.9ĩ}$$

$$\delta\tilde{\beta}+\tilde{\delta}\beta=-\mu\rho-4\beta\tilde{\beta}, \tag{2.9j}$$

$$\tilde{\delta}\beta+\delta\tilde{\beta}=-\tilde{\mu}\rho-4\beta\tilde{\beta}+\tilde{\Psi}_2, \tag{2.9j̃}$$

$$\tilde{\delta}\mu=0, \tag{2.9k}$$

$$\tilde{\delta}\tilde{\lambda}-\delta\tilde{\mu}=-4\tilde{\beta}\tilde{\lambda}-\tilde{\Psi}_3, \tag{2.9k̃}$$

$$\delta\nu-\Delta\mu=\mu^2+(\gamma+\tilde{\gamma})\mu-2\beta\nu, \tag{2.9l}$$

$$\tilde{\delta}\tilde{\nu}-\Delta\tilde{\mu}=\tilde{\mu}^2+(\gamma+\tilde{\gamma})\tilde{\mu}-2\tilde{\beta}\tilde{\nu}, \tag{2.9l̃}$$

$$\delta\gamma-\Delta\beta=-\sigma\nu+(\mu-\gamma+\tilde{\gamma})\beta-\tilde{\beta}\tilde{\lambda}, \tag{2.9m}$$

$$\tilde{\delta}\tilde{\gamma}-\Delta\tilde{\beta}=(\tilde{\mu}-\tilde{\gamma}+\gamma)\tilde{\beta}, \tag{2.9m̃}$$

$$\Delta\sigma=-\mu\sigma-\rho\tilde{\lambda}+(3\gamma-\tilde{\gamma})\sigma, \tag{2.9n}$$

$$\Delta\rho=(\gamma+\tilde{\gamma}-\tilde{\mu})\rho, \tag{2.9o}$$

$$\Delta\rho=(\tilde{\gamma}+\gamma-\mu)\rho-\tilde{\Psi}_2, \tag{2.9õ}$$

$$\Delta\tilde{\beta}+\tilde{\delta}\gamma=-\rho\nu-(\tilde{\mu}-\tilde{\gamma})\tilde{\beta}-\tilde{\beta}\gamma, \tag{2.9p}$$

$$\Delta\beta+\delta\tilde{\gamma}=-\rho\tilde{\nu}+\tilde{\beta}\tilde{\lambda}-(\mu-\gamma)\beta-\beta\tilde{\gamma}-\tilde{\Psi}_3. \tag{2.9p̃}$$

Bianchi identities

$$D\tilde{\Psi}_2=3\rho\tilde{\Psi}_2, \tag{2.10ã}$$

$$D\tilde{\Psi}_3-\delta\tilde{\Psi}_2=2\rho\tilde{\Psi}_3, \tag{2.10b̃}$$

$$D\tilde{\Psi}_4-\delta\tilde{\Psi}_3=\rho\tilde{\Psi}_4-2\beta\Psi\quad 2\tilde{\beta}\tilde{\Psi}_3, \tag{2.10c̃}$$

$$\tilde{\delta}\tilde{\Psi}_2=0, \tag{2.10d̃}$$

$$\Delta\tilde{\Psi}_2-\tilde{\delta}\tilde{\Psi}_3=-3\tilde{\mu}\tilde{\Psi}_2+2\tilde{\beta}\tilde{\Psi}_3, \tag{2.10ẽ}$$

$$\Delta\tilde{\Psi}_3-\tilde{\delta}\tilde{\Psi}_4=3\tilde{\nu}\tilde{\Psi}_2-2(\tilde{\gamma}+2\tilde{\mu})\tilde{\Psi}_3+4\tilde{\beta}\tilde{\Psi}_4. \tag{2.10f̃}$$

Metric equations

$$D\omega=\rho\omega+\sigma\tilde{\omega}, \tag{2.11a}$$

$$D\tilde{\omega}=\rho\tilde{\omega}, \tag{2.11ã}$$

$$D\xi^k=\rho\xi^k+\sigma\xi^k, \tag{2.11b}$$

$$D\tilde{\xi}^k=\rho\tilde{\xi}^k, \tag{2.11b̃}$$

$$DU=-(\gamma+\tilde{\gamma}), \tag{2.11c}$$

$$DX^k=0, \tag{2.11d}$$

$$\delta U-\Delta\omega=(\mu-\gamma+\tilde{\gamma})\omega+\tilde{\lambda}\tilde{\omega}-\tilde{\nu}, \tag{2.11e}$$

$$\tilde{\delta}U-\Delta\tilde{\omega}=(\tilde{\mu}-\tilde{\gamma}+\gamma)\tilde{\omega}-\nu, \tag{2.11ẽ}$$

$$\delta X^k-\Delta\xi^k=(\mu-\gamma+\tilde{\gamma})\xi^k+\tilde{\lambda}\tilde{\xi}^k, \tag{2.11f}$$

$$\tilde{\delta}X^k-\Delta\tilde{\xi}^k=(\tilde{\mu}-\tilde{\gamma}+\gamma)\tilde{\xi}^k, \tag{2.11f̃}$$

$$\tilde{\delta}\omega-\delta\tilde{\omega}=-2\tilde{\beta}\omega+2\beta\tilde{\omega}+\tilde{\mu}-\mu, \tag{2.11g}$$

$$\tilde{\delta}\xi^k-\delta\tilde{\xi}^k=-2\tilde{\beta}\xi^k+2\beta\tilde{\xi}^k. \tag{2.11h}$$

Though the integration of these equations is straightforward,[8,7] it is rather lengthy and cumbersome. We will thus omit the details and simply present the final results. Along the way we used up nearly all of the coordinate and tetrad freedom to simplify the results of the integrations.

An essential final result is that all variables can be expressed as explicit functions of the four coordinates and four arbitrary complex functions ($\tilde{\Psi}_2^0$, G, A, B) of the two coordinates u, ζ.

The results follow:

Spin coefficients

$$\rho=-1/r, \tag{2.12a}$$

$$\sigma=\sigma^0/r^2, \tag{2.12b}$$

$$\beta=\beta^0/r+\tilde{\beta}^0\sigma^0/r^2, \tag{2.12c}$$

$$\tilde{\beta}=\tilde{\beta}^0/r, \tag{2.12c̃}$$

$$\mu=\mu^0/r, \tag{2.12d}$$

$$\tilde{\mu}=\mu^0/r-\tilde{\Psi}_2^0/r^2, \tag{2.12d̃}$$

$$\nu=\nu^0, \tag{2.12e}$$

$$\tilde{\nu}=\tilde{\nu}^0-\frac{\tilde{\Psi}_3^0}{r}+\frac{p\partial\tilde{\Psi}_2^0/\partial\zeta}{2r^2}-\frac{\omega_0\tilde{\Psi}_2^0}{2r^3}, \tag{2.12ẽ}$$

$$\gamma=\gamma^0, \tag{2.12f}$$

$$\tilde{\gamma}=\gamma^0-\tilde{\Psi}_2^0/2r^2, \tag{2.12f̃}$$

$$\tilde{\lambda}=\tilde{\lambda}^0/r-\mu^0\sigma^0/r^2+\sigma^0\tilde{\Psi}_2^0/2r^3, \tag{2.12g̃}$$

where

$$p=1+G\tilde{\zeta}, \tag{2.13a}$$

$$\sigma^0=-\frac{\tilde{\Psi}_2^0}{2G^2}-\frac{pA}{G}+p^2B, \tag{2.13b}$$

$$\beta^0=-\frac{1}{2}p\frac{\partial\ln p}{\partial\zeta}, \tag{2.13c}$$

$$\tilde{\beta}^0=-\frac{1}{2}p\frac{\partial\ln p}{\partial\tilde{\zeta}}, \tag{2.13c̃}$$

$$\mu^0=-p^2\frac{\partial^2\ln p}{\partial\zeta\partial\tilde{\zeta}}, \tag{2.13d}$$

$$\nu^0=-p\frac{\partial^2\ln p}{\partial u\partial\tilde{\zeta}}, \tag{2.13e}$$

$$\tilde{\nu}^0=-p\frac{\partial^2\ln p}{\partial u\partial\zeta}, \tag{2.13ẽ}$$

$$\gamma^0=-\frac{1}{2}\frac{\partial\ln p}{\partial u}, \tag{2.13f}$$

$$\tilde{\lambda}^0=p\frac{\partial(\sigma^0/p)}{\partial u}, \tag{2.13g̃}$$

and $\tilde{\Psi}_3^0$ and ω^0 are defined in Eqs. (2.15ã) and (2.17), respectively.

Weyl tensor

$$\tilde{\Psi}_2=\frac{\tilde{\Psi}_2^0}{r^3}, \tag{2.14ã}$$

$$\tilde{\Psi}_3=\frac{\tilde{\Psi}_3^0}{r^2}-\frac{p}{r^3}\frac{\partial\tilde{\Psi}_2^0}{\partial\zeta}+\frac{3\omega^0\tilde{\Psi}_2^0}{2r^4}, \tag{2.14b̃}$$

$$\tilde{\Psi}_4=\frac{\tilde{\Psi}_4^0}{r}+\left(2\beta^0\tilde{\Psi}_3^0-p\frac{\partial\tilde{\Psi}_3^0}{\partial\zeta}\right)\frac{1}{r^2}+\left[(\beta^0\sigma^0+\omega^0)\tilde{\Psi}_3^0+\frac{1}{2}p\sigma^0\frac{\partial\tilde{\Psi}_3^0}{\partial\tilde{\zeta}}\right.$$
$$\left.+\frac{1}{2}p\frac{\partial}{\partial\zeta}\left(p\frac{\partial\tilde{\Psi}_2^0}{\partial\zeta}\right)-p\beta^0\frac{\partial\tilde{\Psi}_2^0}{\partial\zeta}+\frac{3}{2}\tilde{\lambda}^0\tilde{\Psi}_2^0\right]\frac{1}{r^3}$$
$$+\left[\left(\beta^0\omega^0-\mu^0\sigma^0-\frac{1}{2}p\frac{\partial\omega^0}{\partial\zeta}\right)\tilde{\Psi}_2^0-\frac{3}{2}p\omega^0\frac{\partial\tilde{\Psi}_2^0}{\partial\zeta}\right]\frac{1}{r^4}+\frac{3(\omega^0)^2\tilde{\Psi}_2^0}{2r^5}, \tag{2.14c̃}$$

where

$$\tilde{\Psi}_3^0=p\frac{\partial\mu^0}{\partial\zeta}-p^3\frac{\partial}{\partial\tilde{\zeta}}\left(\frac{\tilde{\lambda}^0}{p^2}\right), \tag{2.15ã}$$

$$\tilde{\Psi}_4^0=\frac{\partial(p\tilde{\nu}^0)}{\partial\zeta}-p^2\frac{\partial}{\partial u}\left(\frac{\tilde{\lambda}^0}{p^2}\right). \tag{2.15b̃}$$

Metric variables

$$X^k=0, \tag{2.16a}$$

$$U=-2\gamma^0 r+\mu^0-\tilde{\Psi}_2^0/2r, \tag{2.16b}$$

$$\omega=\frac{\omega^0}{r}, \tag{2.16c}$$

$$\tilde{\omega}=0, \tag{2.16c̃}$$

$$\xi^k=(p/r)\delta_2^k-(p\sigma^0/r^2)\delta_3^k, \tag{2.16d}$$

$$\tilde{\xi}^k=(p/r)\delta_3^k, \tag{2.16d̃}$$

where

$$\omega^0=\frac{\tilde{\Psi}_2^0}{G}+pA. \tag{2.17}$$

In terms of these quantities, the metric takes the form

$$g^{\mu\nu}=\begin{pmatrix}0&1&0&0\\1&2U&0&-\tilde{\xi}^3\omega\\0&0&0&-\xi^2\tilde{\xi}^3\\0&-\tilde{\xi}^3\omega&-\xi^2\tilde{\xi}^3&-2\xi^3\tilde{\xi}^3\end{pmatrix}, \tag{2.18a}$$

$$g_{\mu\nu}=\begin{pmatrix}-2U&1&-\omega/\xi^2&0\\1&0&0&0\\-\omega/\xi^2&0&2\xi^3/(\xi^2)^2\tilde{\xi}^3&-1/\xi^2\tilde{\xi}^3\\0&0&-1/\xi^2\tilde{\xi}^3&0\end{pmatrix}. \tag{2.18b}$$

In obtaining this solution, we have reduced the remaining coordinate freedom to transformations of the form

$$u^*=(Y')^{1/2}u+f(\zeta), \tag{2.19a}$$

$$r^*=(Y')^{-1/2}r, \tag{2.19b}$$

$$\zeta^*=Y(\zeta), \tag{2.19c}$$

$$\tilde{\zeta}^*=\tilde{\zeta}-\frac{p^2h(u,\zeta)}{r+hGp}, \tag{2.19c̃}$$

where f and Y are arbitrary functions of ζ and

$$Y'=\frac{dY}{d\zeta},\quad h(u,\zeta)=\frac{\partial u^*/\partial\zeta}{\partial u^*/\partial u}. \tag{2.20}$$

III. TYPES III, N, AND D

In the previous section we obtained the general algebraically special H-space metric with nonvanishing divergence. We now show how further specialization leads to types III, N and D.

The condition for a type III metric,

$$\tilde{\Psi}_0=\tilde{\Psi}_1=\tilde{\Psi}_2=0,\quad \tilde{\Psi}_3\neq 0, \tag{3.1}$$

leads immediately to the restriction that

$$\tilde{\Psi}_2^0(u,\zeta)=0. \tag{3.2}$$

Thus the solution is expressed in terms of the three functions G, A, B of u, ζ.

Similarly, for a type N metric we must have $\tilde{\Psi}_0=\tilde{\Psi}_1=\tilde{\Psi}_2=\tilde{\Psi}_3=0$, $\tilde{\Psi}_4\neq 0$, which leads to the conditions $\tilde{\Psi}_2^0=\tilde{\Psi}_3^0=0$. It follows then from Eqs. (2.15) and (2.13) that

$$B=-\left(\frac{\partial G}{\partial u}\right)^{-1}\left[\frac{\partial^2 G}{\partial\zeta^2}+G\frac{\partial}{\partial u}\left(\frac{A}{G}\right)\right]. \tag{3.3}$$

We thus see that the general N solution depends only on the two functions G and A.

H-spaces of Petrov type D are those in which it is possible to choose a tetrad such that

$$\Psi_0=\Psi_1=\Psi_2=\Psi_3=\Psi_4=0, \tag{3.4a}$$

$$\tilde{\Psi}_0=\tilde{\Psi}_1=\tilde{\Psi}_3=\tilde{\Psi}_4=0. \tag{3.4ã}$$

This choice of tetrad, however, is not necessarily the one in which the results of Sec. II are expressed. In order to write the type D solutions in such a tetrad system, we start with a tetrad satisfying Eqs. (3.4) and transform to one satisfying the conditions imposed in Sec. II. When this is done, we find first of all that G is a function only of ζ. If G is not constant, then we may always make the simple choices

$$G=\zeta, \tag{3.5a}$$

$$A=B=0, \tag{3.5b}$$

by using the available coordinate freedom. It then follows that

$$\tilde{\Psi}_2^0=\phi_0, \tag{3.5c}$$

where ϕ_0 is an arbitrary constant. The remaining coordinate freedom is then given by Eqs. (2.19) with $Y=\zeta$ and f an arbitrary constant. If G is constant, then two further cases arise according to whether or not $\tilde{\Psi}_2^0$ is constant. In the first of these cases, we may use the available coordinate freedom to obtain

$$G=G_0, \tag{3.6a}$$

$$\tilde{\Psi}_2^0=1, \tag{3.6b}$$

$$A=A_0, \tag{3.6c}$$

$$B=0, \tag{3.6d}$$

where G_0 and A_0 are arbitrary constants. The remaining coordinate freedom is then given by Eqs. (2.19) with $Y=\zeta+a$ and $f=b\zeta+c$, where a, b, and c are arbitrary constants. In the remaining case, we obtain

$$G=G_0, \tag{3.7a}$$

$$\tilde{\Psi}_2^0=\phi_0/u^3, \tag{3.7b}$$

$$A=B=0, \tag{3.7c}$$

where G_0 and ϕ_0 are arbitrary constants. The remaining coordinate freedom is then given by Eqs. (2.19) with $f=0$ and

$$Y=\frac{c-b^2+ac\zeta}{a(1+a\zeta)},$$

where a, b, and c are arbitrary constants. This somewhat complicated way of choosing the constants in Y was made to simplify the resulting forms of Eqs. (2.19a) and (2.19b), and to facilitate the reduction to the identity transformation (which arises from first setting $b=c=1$ and then taking $a=0$). The complete type D solutions are thus obtained by using Eqs. (3.5), (3.6), or (3.7) in the results of Sec. II. We note that these solutions are completely determined by either one arbitrary constant, in the case when G is not constant, or two arbitrary constants, in the remaining cases.

IV. CONCLUSION

We have seen that with relative ease it has been possible to integrate the $\mathcal{H}$-space spin-coefficient equations to obtain all diverging algebraically special solutions. It remains an open question as to whether the general $\mathcal{H}$-space metric can be found by direct integration, though it is certain that many special cases can be found.[9]

Of possibly greater interest would be to find a method of going from $\mathcal{H}$-space to physical space. We know that an asymptotically flat space—time determines (by means of its radiation field) an $\mathcal{H}$-space (i.e., the space of complex asymptotically shear-free null cones). The inverse question would be how one determines the radiation field of the physical space from a known $\mathcal{H}$-space. This question is actively being explored.

One would further like to know what physical meaning one can give to a particular $\mathcal{H}$-space metric if the $\mathcal{H}$-space is interpreted as a Penrose nonlinear graviton. We appear to be at an impasse on this question.

*This paper incorporates some of the results contained in a Ph.D. dissertation submitted by C.W. Fette to the University of Pittsburgh.
†Research supported in part by a grant from the National Science Foundation.

[1]E.T. Newman, Seventh International Conference on Gravitation and Relativity, Tel-Aviv (1974).
[2]E.T. Newman, Riddle of Gravitation Symposium, Syracuse (1975).
[3]R. Penrose, Riddle of Gravitation Symposium, Syracuse (1975).
[4]R.O. Hansen, E.T. Newman, and R. Penrose, in preparation.
[5]E.T. Newman and J. Winicour, J. Math. Phys. **15**, 1113 (1974).
[6]E.T. Newman and R. Penrose, J. Math. Phys. **3**, 566 (1962).
[7]Further details may be found in C.W. Fette's Ph.D. dissertation, University of Pittsburgh (1975).
[8]E.T. Newman and L.A. Tamburino, J. Math. Phys. **3**, 902 (1962).
[9]J.F. Flebański, J. Math. Phys. **16**, 2395 (1975).

A class of states on the boson–fermion algebra

F. Rocca,* P. N. M. Sisson,† and A. Verbeure

Instituut voor Theoretische Fysica, University of Leuven, Celestijnenlaan 200 D, B-3030 Heverlee, Belgium
(Received 1 December 1975)

On the tensor product C^*-algebra of bosons and fermions a class of states determined by the two-point functions is proved to exist. It is indicated how they induce a class of states on the CCR algebra which are not quasifree, but determined by their two- and four-point functions.

1. INTRODUCTION

The treatment of models which have Hamiltonians at most quadratic in the field variables for bosons or fermions individually leads naturally to the study of quasifree states.[1–4] The essential property of these states is that all correlation functions are expressible in terms of the one- and two-point functions. For the individual systems this is achieved by having all n-point truncated correlation functions zero for $n \geq 3$. Alternatively one can compute the higher correlation functions from the explicit representations for these quasifree states, which are well known.

It is natural, therefore, to consider states on the product (boson × fermion) algebra and to seek states for which all correlation functions are expressed in terms of the one- and two-point functions. Furthermore one may conjecture that such states describe systems for which the Hamiltonian is at most quadratic in either field variable. One such system is the Dicke Maser model.[5,6] Because of the mixed statistics in the product algebra it is not clear how to define truncated correlation functions and consequently we take the alternative procedure and construct a representation of a state on the product algebra with given one- and two-point functions. As a starting point we consider for simplicity the case of one degree of freedom for both bosons and fermions.

The explicit construction of the representation is given in Theorem III. 1. We then consider the restriction of the corresponding state to the CCR subalgebra, and study its properties. We observe that it is not quasifree, nor gauge-invariant despite requiring gauge invariance in the one- and two-point functions.

Finally, we would like to stress that the representations considered in Theorem III. 1 below, are to our knowledge the first explicit ones on the product algebra which yield states which are not simply product states.

2. NOTATION

Let $\mathfrak{A} = \Delta(\mathbb{R}^2, \sigma)$ denote the CCR C^*-algebra for one degree of freedom,[7] and $\mathfrak{B} = \mathfrak{A}(\mathbb{R}^2, s)$ the CAR C^*-algebra,[2] also for one degree of freedom. $s(\cdot,\cdot)$ and $\sigma(\cdot,\cdot) = -s(J\cdot,\cdot)$ are the real imaginary parts of the usual complex inner product on $\mathbb{C} = \mathbb{R} \oplus J\mathbb{R}$.

$\mathfrak{A}$ is generated by $\{W(x) : x \in \mathbb{R}^2\}$ satisfying the commutation relation

$$W(x)\,W(y) = \exp\{-i\sigma(x,y)\}\,W(x+y),$$

and $\mathfrak{B}$ by $\{B(x) : x \in \mathbb{R}^2\}$ satisfying

$$B(x)^* = B(x),$$
$$B(\lambda x + \mu y) = \lambda B(x) + \mu B(y), \quad \lambda, \mu \in \mathbb{R},$$
$$B(x)^2 = s(x,x).$$

Let $\mathfrak{C}$ denote the tensor product C^*-algebra $\mathfrak{A} \otimes \mathfrak{B}$, which is generated by $\{W(x)B(y) : x, y \in \mathbb{R}^2\}$.

We introduce the Fock representation of $\mathfrak{A}$ and $\mathfrak{B}$, and denote the GNS triples by $(\mathcal{H}, \Pi_B, \Omega_B)$ and $(\mathbb{C}^2, \Pi_F, \Omega_F)$ respectively. The corresponding creation a^*, b^* and annihilation a, b operators are defined by

$$\Pi_B(W(e)) = \exp(a^* - a),$$

$$\Pi_F(B(e) - iB(Je)) = 2b^*, \text{ where } e = (1,0) \in \mathbb{R}^2,$$

and satisfy

$$a\Omega_B = b\Omega_F = 0.$$

We will only be interested in regular representations of states on $\mathfrak{C}$ having the two-point functions

$$\omega(aa^*) = \rho_1, \tag{1}$$

$$\omega(bb^*) = \rho_2, \tag{2}$$

$$\omega(a^*b) = \eta, \tag{3}$$

and all other one- and two-point functions zero.

The positivity of the state ω implies necessarily that

$$\rho_1 \geq 1, \quad 1 \geq \rho_2 \geq 0,$$
$$(\rho_1 - 1)(1 - \rho_2) \geq |\eta|^2, \quad \rho_1\rho_2 \geq |\eta|^2. \tag{4}$$

For the individual boson or fermion systems the necessary positivity conditions arising from the one- and two-point functions are also sufficient in defining the quasifree state. However it is not clear in the product case that conditions (4) are sufficient to give rise to positivity of a state on the full algebra $\mathfrak{C}$.

3. REPRESENTATION

Let $\mathfrak{H} = \mathcal{H} \otimes \mathcal{H} \otimes \mathbb{C}^2 \otimes \mathbb{C}^2$ be the representation space and $\Omega = \Omega_B \otimes \Omega_B \otimes \Omega_F \otimes \Omega_F$ the cyclic vector. Define

$$\Pi_1(a) = (\alpha a + \beta a^*) \otimes 1 + 1 \otimes \gamma(a^* + a) + \sigma 1 \otimes 1,$$

$$\Pi_1'(a) = \gamma(a^* + a) \otimes 1 + 1 \otimes (\alpha a + \beta a^*) + \sigma' 1 \otimes 1,$$

$$\Pi_2(b) = \sqrt{1-\rho_2}\, b \otimes 1 + \sqrt{\rho_2}\, \theta \otimes b^*,$$

$$\Pi_2'(b) = \sqrt{\rho_2}\theta b^* \otimes \theta + \sqrt{1-\rho_2}\, 1 \otimes b\theta,$$

where

$$\theta = 2b^*b - 1, \quad \alpha = \bar{\alpha}, \quad \beta = \bar{\beta},$$
$$\bar{\gamma} = -\gamma, \quad \alpha^2 - \beta^2 = 1, \quad \alpha, \beta, \gamma \in \mathbb{C}. \tag{5}$$

It is easily checked that Π_1, Π_1' are $*$-representations of $\mathfrak{A}$ on $H \otimes H$ such that $[\Pi_1(a), \Pi_1'(a)] = [\Pi_1(a^*), \Pi_1'(a)] = 0$, and Π_2, Π_2' are $*$-representations of $\mathfrak{B}$ on $\mathbb{C}^2 \otimes \mathbb{C}^2$ such that $[\Pi_2(b), \Pi_2'(b)] = [\Pi_2(b^*), \Pi_2'(b)] = 0$. Let

$$\Pi(a) = \Pi_1(a) \otimes \Pi_2'(B(e)),$$

$$\Pi(b) = \Pi_1'(W(z)) \otimes \Pi_2(b), \quad z \in \mathbb{R} \oplus J\mathbb{R}.$$

Theorem III. 1: For fixed ρ_1, ρ_2 there exists $\eta_0 \neq 0$ such that whenever $|\eta| \leq \eta_0$ there exist constants α, β, γ, z, σ, σ' such that the vector state

$$\omega_\Pi(x) = (\Omega, \Pi(x)\Omega), \quad x \in \mathfrak{C}$$

satisfies Eqs. (1)–(3).

Proof: The definitions of Π_i, Π_i' $(i = 1, 2)$, and Eqs. (5) imply that π is a $*$-representation of $\mathfrak{C}$. It is a straightforward matter to compute the two-point functions of ω_Π. They yield:

$$\alpha^2 + |\gamma|^2 + |\sigma|^2 = \rho_1 \quad \text{[from (1)]} \tag{6}$$

$$\sqrt{\rho_2(1-\rho_2)}\ (\alpha\bar{u} - \gamma v - \sigma)$$
$$\times \exp\{-\tfrac{1}{2}(|u|^2 + |v|^2) + w\} = \eta \quad \text{[from (3)]}, \tag{7}$$

$$\beta\bar{u} - \bar{\gamma}v - \sigma = 0 \quad [\text{from } \omega_\Pi(a^*b^*) = 0], \tag{8}$$

$$\alpha\beta + \gamma^2 + \sigma^2 = 0 \quad [\text{from } \omega_\Pi(a^2) = 0], \tag{9}$$

where

$$u = \bar{z}\bar{\gamma} - z\gamma, \quad v = \bar{z}\alpha - z\beta, \quad w = \bar{z}\bar{\sigma}' - z\sigma'.$$

The other one- and two-point functions are automatically satisfied. We put $z = \bar{z}$, $\gamma = i\gamma_0$, and then obtain:

$$u = -2iz\gamma_0, \quad v = z(\alpha - \beta), \quad w = z(\bar{\sigma}' - \sigma'),$$

$$\sigma = -i\gamma_0 z(\alpha + \beta) \equiv i\sigma_0.$$

Equations (6)–(9) become

$$\alpha = \rho_1/\sqrt{2\rho_1 - 1}, \quad \beta = (\rho_1 - 1)/\sqrt{2\rho_1 - 1},$$

$$z(\bar{\sigma}' - \sigma') = i(\pi/2 + \arg\eta), \quad \gamma_0^2 = \alpha\beta - \sigma_0^2,$$

$$z^2 = \sigma_0^2(\alpha - \beta)^2/(\alpha\beta - \sigma_0^2),$$

and finally

$$-\log\lambda^2 + \log\sigma_0^2 = \sigma_0^2(1 - B\sigma_0^2)/(A - \sigma_0^2), \tag{10}$$

where

$$\lambda = |\eta|/2\sqrt{\rho_2(1-\rho_2)}, \quad A = \rho_1(\rho_1 - 1)/(2\rho_1 - 1),$$
$$B = 4/(2\rho_1 - 1).$$

Let $f_\lambda(x) = \log x^2 - \log\lambda^2$, $g(x) = x^2(1 - Bx^2)/(A - x^2)$. f_λ and g are continuous functions on $[0, A)$ and g is independent of λ (see Fig. 1). Consequently $\exists \lambda_0$ such that whenever $\lambda < \lambda_0$ there is a solution to $f_\lambda(x) = g(x)$. Let $\eta_0 = 2\lambda_0\sqrt{\rho_2(1-\rho_2)}$, then for $|\eta| < \eta_0$ there are two solutions to (10) and for $|\eta| = \eta_0$ there exists one solution. Hence the theorem is proved.

Remark III. 2: In order to obtain an idea of how small η_0 has to be, we show now that for $\lambda_0^2 \geq (\rho_1 - 1)/4$ there is no solution to (10).

It is enough to show it for $\lambda_0^2 = (\rho_1 - 1)/4$. It is straightforward to see that in this case one has $g(x) \geq x^2/\lambda_0^2 - 1 \geq f_{\lambda_0}(x)$, using the inequality $\log x \leq x - 1$ for $x \geq 1$. In words, a representation of the type considered in Theorem III. 1 exists when the correlation function $\eta = \omega(a^*b)$ is small enough.

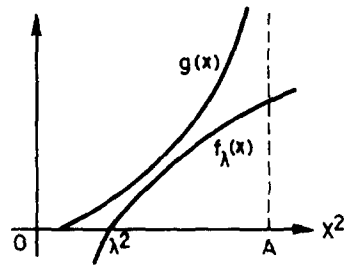

FIG. 1.

Remark III. 3: We consider the generating function $\mu_\Pi(x) = \omega_\Pi(W(x) \otimes 1)$, where $W(x) = \exp(\bar{x}a^* - xa)$,

$$\Pi(\bar{x}a^* - xa) = R \otimes \Pi_2'(B(e)),$$

$$R = (\bar{\lambda}a^* - \lambda a) \otimes 1 + 1 \otimes \mu(a^* + a) + \nu(1 \otimes 1),$$

$$\lambda = x\alpha - \bar{x}\beta, \quad \mu = \bar{x}\bar{\gamma} - x\gamma, \quad \nu = \bar{x}\bar{\sigma} - x\sigma.$$

So

$$[\Pi(\bar{x}a^* - xa)]^{2k} = R^{2k} \otimes 1 \otimes 1,$$

$$[\Pi(\bar{x}a^* - xa)]^{2k+1} = R^{2k+1} \otimes \Pi_2'(B(e)).$$

Hence

$$\Pi(W(x)) = \cosh R \otimes 1 \otimes 1 + \sinh R \otimes \Pi_2'(B(e)),$$

and after simple calculation

$$\mu_\Pi(x) = \exp[-\tfrac{1}{2}(2\rho_1 - 1)|x|^2]$$
$$\times \exp[+\tfrac{1}{2}\sigma_0^2(x + \bar{x})^2] \cos[\sigma_0(x + \bar{x})]. \tag{11}$$

μ_Π is not the generating function of a quasifree state, and from its form it is clearly not gauge invariant. This can be checked by computing the four point function $\omega_\Pi(a^{*3}a) = 2\sigma_0^4$.

Let $\omega_\Pi^{\mathfrak{A}}$ be the restriction of the state ω_Π to the C^*-subalgebra $\mathfrak{A}$ of $\mathfrak{C}$. The gauge invariant part of $\omega_\Pi^{\mathfrak{A}}$ is given by the generating function

$$\bar{\mu}_\Pi(x) = \int_0^{2\pi} \mu_\Pi(e^{i\alpha}x)\frac{d\alpha}{2\pi} = \exp[-\tfrac{1}{2}(2\rho_1 - 1)|x|^2]$$
$$\times \int_0^{2\pi} \exp[\tfrac{1}{2}\sigma_0^2(x\exp(i\alpha) + \bar{x}\exp(-i\alpha))^2]$$
$$\times \cos(\sigma_0(x\exp(i\alpha) + \bar{x}\exp(-i\alpha)))\frac{d\alpha}{2\pi}.$$

Remark III. 4: A quasifree state over $\Delta(R^2, \sigma)$ is defined completely in terms of the one- and two-point correlation functions. The state $\omega_\Pi^{\mathfrak{A}}$ is also described by two parameters ρ_1, σ_0 where

$$\rho_1 = \omega_\Pi(aa^*), \quad 2\sigma_0^4 = \omega_\Pi(a^{*3}a),$$

and so is a state determined by its two- and four-point functions. Indeed from (11) one immediately sees that $\mu_\Pi = \frac{1}{2}(\nu_\Pi + \nu_\Pi^*)$, where ν_Π is the generating function of a quasifree state, given by

$$\nu_\Pi(x) = \exp[-\tfrac{1}{2}(2\rho_1 - 1)|x|^2 + \tfrac{1}{2}\sigma_0^2(x + \bar{x})^2 + i\sigma_0(x + \bar{x})].$$

ACKNOWLEDGMENT

One of us (F. R.) thanks the members of the Institut voor Theoretische Fysika, University of Leuven, for their kind hospitality.

*On leave from: Institut de Physique Théorique, Université de Nice, France.
†Derde Cyclus Fellow, K.U. Leuven, Belgium.

[1]D.W. Robinson, Commun. Math. Phys. **1**, 159 (1965).
[2]E. Balslev, J. Manuceau, and A. Verbeure, Commun. Math. Phys. **8**, 315 (1968).
[3]J. Manuceau and A. Verbeure, Commun. Math. Phys. **9**, 293 (1969).
[4]R. T. Powers, Ph. D. Thesis, Princeton (1968).
[5]K. Hepp and E. Lieb, Ann. Phys. **76**, 360 (1973).
[6]M. Fannes, P. Sisson, A. Verbeure, and J. Wolfe, Ann. Phys. to appear.
[7]J. Manuceau, M. Sirugue, D. Testard, and A. Verbeure, Commun. Math. Phys. **32**, 231 (1973).

U(5) ⊃ O(5) ⊃ O(3) and the exact solution for the problem of quadrupole vibrations of the nucleus

E. Chacón* and M. Moshinsky*†

Instituto de Física, Universidad de México (UNAM) México, D.F., Mexico

R. T. Sharp‡

Centre de Recherches Mathématiques, Université de Montréal, Canada
(Received 3 November 1975)

Over twenty years ago A. Bohr discussed the quantum mechanical problem of the quadrupole vibrations in the liquid drop model of the nucleus. States of definite angular momentum L could not be obtained exactly except when $L = 0,3$. In the present paper we indicate how we can determine states for arbitrary angular momentum L and definite number of quanta ν in terms of polynomials of the creation operators characterized by irreducible representation (IR) of the chain of groups U(5)⊃O(3). We furthermore characterize the states by a definite IR λ of O(5) by replacing the creation operators by traceless ones. These states are fully determined by an extra label μ that gives the number of triplets of traceless creation operators coupled to angular momentum zero. We show then how all the wavefunctions of the problem discussed by Bohr can be obtained in a recursive fashion and briefly discuss some of their applications.

1. INTRODUCTION

Over twenty years ago Bohr[1] discussed the quadrupole vibrations of the liquid drop in the quantum mechanical picture. This problem provided the basis for the introduction of the collective degrees of freedom in the phenomena of nuclear structure, that was so fruitful in the development of nuclear physics.

In his analysis Bohr and his collaborators[1,2,3] went into what is known as the strong coupling picture to construct the required states. These wavefunctions can then be characterized by the number of quanta ν, seniority λ, angular momentum L and projection M. The states of lowest even angular momentum, i.e., $L = 0$ and arbitrary ν, λ were already available in the papers mentioned and those of lowest odd angular momentum, i.e., $L = 3$ were obtained shortly afterwards.[4] Yet, as far as we know, no systematic procedure is given in the literature to determine the states for arbitrary L, as well as to characterize the missing label (which we shall denote by μ) in such a way that we get a fully defined and complete set of states.

We plan to carry out this program in the present paper. We shall start in the next section by briefly reviewing Bohr's analysis[1] stressing in particular the part of the problem, connected with the γ vibrations, that remained to be solved.

We then formulate in Sec. 3 the group theoretical structure of the problem and indicate how it can lead to determination of states characterized by a definite irreducible representation ν of U(5) and L of its subgroup O(3). The arguments, though group theoretical, use only concepts related with the theory of angular momenta and elementary algebra.[5] The states are obtained as polynomials in the creation operators and are furthermore characterized by two other missing labels.

Finally in Sec. 4 we introduce the concept of the traceless creation operator recently discussed by Lohe[6] starting from considerations introduced by Vilenkin.[7] Replacing the creation operators mentioned in the previous paragraph with traceless ones, we automatically get states characterized by the IR ν of U(5), λ of O(5), L of O(3), and M of O(2). The missing label is not related with the eigenvalue of the Casimir operator of any group but, in analogy with the concept of seniority,[8] with the number μ of triplets of traceless creation operators that are coupled to angular momentum 0. Expressing the traceless creation operators in terms of the variables appearing in the strong coupling picture we can, in a recursive fashion, get all states from those of $L = 0$ or $L = 3$ that were mentioned previously.

In conclusion we indicate some possible applications of the states that are explicitly derived in the present paper.

2. QUADRUPOLE VIBRATIONS OF THE LIQUID DROP IN THE STRONG COUPLING PICTURE

If the motion of the liquid drop is restricted to the quadrupole type the surface of the nucleus is described by the equation

$$R = R_0\left(1 + \sum_m \alpha^m Y_{2m}(\theta, \phi)\right), \tag{2.1}$$

where R_0 is the spherical radius in the absence of deformation, $Y_{lm}(\theta, \phi)$ the spherical harmonics and α^m, $m = 2, 1, 0, -1, -2$, the contravariant form of the generalized coordinates describing the collective motion.

As usual, the covariant form α_m of these collective coordinates is given by

$$\alpha_m = (-)^m \alpha^{-m}, \tag{2.2}$$

and they are basis for an IR of O(3) associated with $l = 2$. Introducing now a_m as the corresponding generalized coordinates in the frame fixed in the body along the principal axes we have the relation[1,9]

$$\alpha_m = \sum_{m'} D^{2*}_{mm'}(\vartheta_i)\, a_{m'}, \tag{2.3}$$

where ϑ_i, $i=1,2,3$ are the Euler angles, and because the inertia tensor becomes diagonal in this frame we have[1,9]

$$a_2 = a_{-2} \equiv (1/\sqrt{2})\beta \sin\gamma, \quad a_1 = a_{-1} = 0, \quad a_0 \equiv \beta\cos\gamma. \tag{2.4}$$

Up to second-order the classical Lagrangian for the motion is

$$L = \tfrac{1}{2}B_2 \sum_m \dot{\alpha}_m \dot{\alpha}^m - \tfrac{1}{2}C_2 \sum_m \alpha^m \alpha_m, \tag{2.5}$$

where the parameters B_2 and C_2 are related with the density, surface tension, and charge of the liquid drop.[10] In the present paper we shall take units in which

$$\hbar = B_2 = C_2 = 1. \tag{2.6}$$

The covariant momenta are then given by

$$\pi_m = \frac{\partial L}{\partial \dot{\alpha}^m} = \dot{\alpha}_m = \sum_{m'} \dot{D}^{2*}_{mm'}(\vartheta_i)\, a_{m'} + \sum_{m'} D^{2*}_{mm'}(\vartheta_i)\, \dot{a}_{m'}. \tag{2.7}$$

The time derivative of Wigner's $D^2(\vartheta_i)$ function is discussed by Eisenberg and Greiner[11] and using this result, as well as (2.4), we can write

$$\pi_m = (\alpha_m/\beta)\, P_\beta + \beta^{-1}\Delta_m, \tag{2.8}$$

where

$$(\alpha_m/\beta) = (1/\sqrt{2})[D^{2*}_{m2}(\vartheta_i) + D^{2*}_{m-2}(\vartheta_i)]\sin\gamma + D^{2*}_{m0}(\vartheta_i)\cos\gamma, \tag{2.9a}$$

$$\Delta_m = i \sum_{m''k} D^{2*}_{mm''}(\vartheta_i)\, \mathcal{J}_k^{-1} \{[\langle 2m'' | L_k | 22\rangle^* + \langle 2m'' | L_k | 2-2\rangle^*](1/\sqrt{2})\,\beta^2 \sin\gamma + \langle 2m'' | L_k | 20\rangle^* \beta^2 \cos\gamma\} L'_k + \{[D^{2*}_{m2}(\vartheta_i) + D^{2*}_{m-2}(\vartheta_i)]\cos\gamma - D^{2*}_{m0}(\vartheta_i)\sin\gamma\} P_\gamma. \tag{2.9b}$$

In (2.9b), $\langle 2m'' | L_k | 2m\rangle^*$ is the conjugate of the standard matrix element of the angular momentum operator L_k, $k=1,2,3$ and L'_k are the components of the angular momentum in the frame fixed in the body. In the classical picture

$$P_\beta = \dot{\beta}, \tag{2.10a}$$

$$P_\gamma = \beta^2 \dot{\gamma}, \tag{2.10b}$$

while in the quantum picture

$$P_\beta = \frac{1}{i}\frac{\partial}{\partial\beta}, \tag{2.11a}$$

$$P_\gamma = \frac{1}{i}\frac{\partial}{\partial\gamma}, \tag{2.11b}$$

and the L'_k take the operator form, in terms of Euler angles and their derivatives, discussed in Ref. 11. The principal moments of inertia $\mathcal{J}_k$, $k=1,2,3$, along the principal axes are given by[11]

$$\mathcal{J}_k \equiv \beta^2 I_k, \tag{2.12a}$$

$$I_k = 4\sin^2(\gamma - 2\pi k/3). \tag{2.12b}$$

The quantum mechanical Hamiltonian in the body fixed frame of reference is given by[1,12]

$$H = -\frac{1}{2}\frac{1}{\beta^4}\frac{\partial}{\partial\beta}\beta^4\frac{\partial}{\partial\beta} + \frac{1}{2\beta^2}\Lambda^2 + \tfrac{1}{2}\beta^2, \tag{2.13}$$

where Λ^2 is the Casimir operator of O(5) that has the form[12]

$$\Lambda^2 = -\frac{1}{\sin 3\gamma}\frac{\partial}{\partial\gamma}\sin 3\gamma\frac{\partial}{\partial\gamma} + \sum_{k=1}^{3} I_k^{-1} L_k'^2. \tag{2.14}$$

The eigenvalue of Λ^2 is given by[13]

$$\lambda(\lambda+3), \tag{2.15}$$

where λ is an integer. As the Hamiltonian H is that of a five-dimensional oscillator its eigenvalue is given in terms of the number ν of quanta by

$$\nu + \tfrac{5}{2} = 2n + \lambda + \tfrac{5}{2}, \tag{2.16}$$

where $n = \frac{1}{2}(\nu - \lambda)$ is now the radial quantum number. The eigenvalues associated with the total angular momentum

$$L^2 = \sum_{k=1}^{3} L_k'^2 = \sum_{k=1}^{3} L_k^2, \tag{2.17}$$

and L_3 are respectively

$$L(L+1), M. \tag{2.18}$$

Thus our states can be denoted by the ket[1,12]

$$|\nu\lambda\mu LM\rangle = F_n^\lambda(\beta) \sum_K \phi_K^{\lambda\mu L}(\gamma)[D^{L*}_{MK}(\vartheta_i) + (-)^L D^{L*}_{M-K}(\vartheta_i)] \tag{2.19}$$

where we indicated by μ the missing one of the five quantum numbers required to characterize the state completely.

In (2.19) we note that the dependence on β can be obtained immediately from (2.13) if we replace Λ^2 by $\lambda(\lambda+3)$, and thus we get[14]

$$F_n^\lambda(\beta) = \left(\frac{2(n!)}{\Gamma(n+\lambda+\frac{5}{2})}\right)^{1/2} \beta^\lambda L_n^{\lambda+3/2}(\beta^2)\exp(-\beta^2/2), \tag{2.20}$$

where $L_n^{\lambda+3/2}(\beta^2)$ is a Laguerre polynomial and the function is normalized for the volume element $\beta^4 d\beta$. For the dependence on γ, ϑ_i, we can use the fact that the $D^{L*}_{MK}(\vartheta_i)$ constitute a complete set of functions of the Euler angles. In fact, from the symmetry considerations[1,12] associated with the choice of principal axes, we see that the development must be made in terms of

$$(D^{L*}_{MK} + (-)\, D^{L*}_{M-K}), K = 0, 2, \ldots, L \text{ for } L \text{ even};$$
$$K = 0, 2, \ldots, L-1 \text{ for } L \text{ odd}, \tag{2.21}$$

rather than in terms of the D^{L*}_{MK} themselves. Thus the only remaining part of the state (2.19) to be determined is the one associated with the variable γ. Applying the operator

$$\Lambda^2 |\nu\lambda\mu LM\rangle = \lambda(\lambda+3)\,|\nu\lambda\mu LM\rangle, \tag{2.22}$$

we clearly see that this provides a set of coupled ordinary differential equations for the

$$\phi_K^{\lambda\mu L}(\gamma),$$

where λ, μ, L are given and the K take the values indicated in (2.21). This is the part that has been solved

exactly only for the special cases $L=0$, 3 that we proceed to discuss. [See note at the end of the paper.]

For $L=0$, $K=0$ and the application of the L'_k operators in (2.14) gives 0, so we are led to the equation

$$-\frac{1}{\sin 3\gamma}\frac{d}{d\gamma}\sin 3\gamma\frac{d}{d\gamma}\phi_0^{\lambda\mu\,0}(\gamma)=\lambda(\lambda+3)\,\phi_0^{\lambda\mu\,0}(\gamma). \quad (2.23)$$

A regular solution is possible only if λ is a multiple of 3 and we define μ in this case by requiring that

$$\lambda=3\,\mu, \quad (2.24)$$

so that the $\phi_0^{3\mu,\mu,0}$ is proportional to the Legendre polynomial $P_\mu(\cos 3\gamma)$. Clearly we can limit discussion of the states (2.19) to $\nu=\lambda$, as $\nu=2n+\lambda$ means only introducing the Laguerre polynomial $L_n^{\lambda+3/2}(\beta^2)$ instead of the unity to which it reduces when $n=0$. We can also restrict our analysis to $M=L$ and thus deal only with the states

$$|\lambda\mu L)\equiv|\nu=\lambda,\lambda,\mu,L,M=L\rangle. \quad (2.25)$$

For $L=0$ they have then the[12] form

$$|3\mu,\mu,0)=A_\mu\beta^{3\mu}\exp(-\beta^2/2)\,P_\mu(\cos 3\gamma) \quad (2.26)$$

where A_μ is so far an arbitrary constant which we shall later select conveniently.

Turning now our attention to odd L we note that (2.19) vanishes identically if $L=1$ because then $K=0$ is the only possible value and the square bracket in it is zero. We look then to $L=3$ for which $K=0,2$, but again $K=0$ vanishes, so we get only a single ordinary differential equation for $\phi_2^{\lambda\mu\,3}(\gamma)$ of the form

$$\left[-\frac{1}{\sin 3\gamma}\frac{d}{d\gamma}\sin 3\gamma\frac{d}{d\gamma}+\frac{9}{\sin^2 3\gamma}\right]\phi_2^{\lambda\mu\,3}(\gamma)=\lambda(\lambda+3)\,\phi_2^{\lambda\mu\,3}(\gamma), \quad (2.27)$$

where we used the fact that

$$(7/8\pi^2)\int D^3_{\pm 2}(\vartheta_i)\,L'^2_k D^{3*}_{\pm 2}(\vartheta_i)\,d\Omega=4,\quad k=1,2,3, \quad (2.28)$$

where $d\Omega=\sin\vartheta_2\,d\vartheta_1\,d\vartheta_2\,d\vartheta_3$, and also that[15]

$$\frac{1}{\sin^2\gamma}+\frac{1}{\sin^2(\gamma-2\pi/3)}+\frac{1}{\sin^2(\gamma-4\pi/3)}=\frac{9}{\sin^2 3\gamma}. \quad (2.29)$$

As in the case of $L=0$ a regular solution exists only if $\lambda\equiv 0 \bmod 3$ and for later convenience we write, in the case $L=3$, that

$$\lambda=3\mu+3 \quad (2.30)$$

We have then that $\phi^{3\mu+3,\mu,3}(\gamma)$ will be proportional to the associated Legendre polynomial $P^1_{\mu+1}(\cos 3\gamma)$ and the state $|3\mu+3,\mu,3)$ takes the form

$$|3\mu+3,\mu,3)=B_\mu\beta^{3\mu+3}\exp(-\beta^2/2)\,P^1_{\mu+1}(\cos 3\gamma)\times[D^{3*}_{32}(\vartheta_i)-D^{3*}_{3-2}(\vartheta_i)], \quad (2.31)$$

where B_μ is again an arbitrary constant to be selected later.

We shall now proceed to derive in a systematic fashion the states $|\lambda,\mu,L)$ for arbitrary L starting from $L=0$ if L is even, or $L=3$ if L is odd.

3. STATES CHARACTERIZED BY THE IR OF THE CHAIN OF GROUPS U(5) ⊃ 0 (3)

We shall proceed to obtain the eigenstates of the Hamiltonian (2.13), of angular momentum L and highest projection $M=L$, as polynomials in the covariant creation operators η_m, $m=2,1,0,-1,-2$, defined as

$$\eta_m=(1/\sqrt{2})(\alpha_m-i\pi_m), \quad (3.1)$$

where α_m is given by (2.3), (2.4) and π_m by (2.8) and (2.9), with the latter being understood as a quantum mechanical operator. These polynomials will be applied to the ground state

$$|0\rangle=(\tfrac{8}{3})^{1/2}\pi^{-1/4}\exp(-\beta^2/2). \quad (3.2)$$

The contravariant form of the annihilation operators is

$$\xi^m=(1/\sqrt{2})(\alpha^m+i\pi^m), \quad (3.3)$$

and they satisfy the commutation relation

$$[\xi^m,\eta_{m'}]=\delta^m_{m'}, \quad (3.4a)$$

which implies that ξ^m, when applied to polynomials in η_m, can be interpreted as

$$\xi^m=\frac{\partial}{\partial\eta_m}. \quad (3.4b)$$

The relation between covariant and contravariant components of the same operator is again (2.2).

The number operator has then the form

$$N=\sum_{m=-2}^{2}\eta_m\xi^m=H-\tfrac{5}{2}, \quad (3.5)$$

and the components of angular momentum are given by[16]

$$L_q=\sum_{mm'}\sqrt{6}\,\langle 21mq\,|\,2m'\rangle\,\eta_{m'}\xi^m,\quad q=1,0,-1, \quad (3.6)$$

implying in particular that

$$L_1=-\sum_{m=-2}^{2}[\tfrac{1}{2}(3+m)(2-m)]^{1/2}\,\eta_{m+1}\xi^m, \quad (3.7a)$$

$$L_0=\sum_{m=-2}^{2}m\eta_m\xi^m. \quad (3.7b)$$

We are now in search of the polynomials $P(\eta_m)$ that satisfy the equations

$$NP=\nu P,\quad L_1P=0,\quad L_0P=LP, \quad (3.8)$$

where the ξ^m in the operators is interpreted in the differential form (3.4).

The first equation in (3.8) implies that P is a homogeneous polynomial in the η's and thus we can write it as

$$P(\eta_m)=\eta_2^\nu P'\left(\frac{\eta_1}{\eta_2},\frac{\eta_0}{\eta_2},\frac{\eta_{-1}}{\eta_2},\frac{\eta_{-2}}{\eta_2}\right), \quad (3.9)$$

where P' is an arbitrary polynomial in the variables indicated of degree not exceeding ν. To apply the second equation of (3.8) we shall introduce the following polynomial functions:

$$(2,0)\equiv\sum_{m=-2}^{2}(-)^m\eta_m\eta_{-m} \tag{3.10a}$$

$$(3,3)\equiv-\sqrt{14/3}\sum_{mm'}\sum_{m''m'''}\langle 22m'''m''|33\rangle \times\langle 22mm'|2m'''\rangle\eta_m\eta_{m'}\eta_{m''}, \tag{3.10b}$$

$$(2,2)\equiv\sqrt{7}\sum_{mm'}\langle 22mm'|22\rangle\eta_m\eta_{m'}. \tag{3.10c}$$

The notation is (ν, L) indicating the number of quanta ν, and the maximum projection $M=L$ of the angular momentum L of the polynomials, which correspond to elementary permissible diagrams (epd).[17,18]

From the explicit form of the Wigner coefficients in (3.10) we obtain

$$\frac{(2,0)}{\eta_2^2}=2\frac{\eta_{-2}}{\eta_2}-2\frac{\eta_{-1}}{\eta_2}\frac{\eta_1}{\eta_2}+\frac{\eta_0^2}{\eta_2^2}, \tag{3.11a}$$

$$\frac{(3,3)}{\eta_2^3}=2\frac{\eta_{-1}}{\eta_2}-\sqrt{6}\frac{\eta_1}{\eta_2}\frac{\eta_0}{\eta_2}+\frac{\eta_1^3}{\eta_2^3}, \tag{3.11b}$$

$$\frac{(2,2)}{\eta_2^2}=2\sqrt{2}\frac{\eta_0}{\eta_2}-\sqrt{3}\frac{\eta_1^2}{\eta_2^2}, \tag{3.11c}$$

and therefore we could also write P as

$$P(\eta_m)=\eta_2^{\nu}P''\left(\frac{\eta_1}{\eta_2},\frac{(2,2)}{\eta_2^2},\frac{(3,3)}{\eta_2^3},\frac{(2,0)}{\eta_2^2}\right), \tag{3.12}$$

where P'' is again an arbitrary polynomial of the variables indicated.

We note now that from their definition,

$$L_1\eta_2=L_1(2,0)=L_1(3,3)=L_1(2,2)=0, \qquad L_1\eta_1=-\sqrt{2}\eta_2, \tag{3.13}$$

and thus, as L_1 given by (3.7a) is a first-order differential operator in the η's, we have

$$L_1P=-\sqrt{2}\eta_2^{\nu}\frac{\partial P''}{\partial(\eta_1/\eta_2)}=0 \tag{3.14}$$

which implies that P'' is independent of (η_1/η_2) and therefore we can write

$$P=\sum_{n_1n_2n_3}B_{n_1n_2n_3}(2,0)^{n_1}(3,3)^{n_2}(2,2)^{n_3}\eta_2^{\nu-2n_1-3n_2-2n_3} \tag{3.15}$$

where so far the $B_{n_1n_2n_3}$ are arbitrary constants.

Considering the third equation in (3.8) we see from (3.7b) and (3.15) that it implies

$$3n_2+2n_3+2(\nu-2n_1-3n_2-2n_3)=L, \tag{3.16}$$

from which we obtain

$$n_3=\tfrac{1}{2}(2\nu-L-3n_2)-2n_1. \tag{3.17}$$

Thus the polynomial that satisfies the Eqs. (3.8) has the form

$$P(\eta_m)=\sum_{n_1n_2}B'_{n_1n_2}\eta_2^{L-\nu+2n_1}(2,0)^{n_1}(3,3)^{n_2}\times(2,2)^{(2\nu-L-3n_2)/2-2n_1}. \tag{3.18}$$

We note from (3.18) that for P to be a polynomial, n_2 must be even (odd) if $2\nu-L$, and thus also L, is even (odd). At first sight the different polynomials seem to be given by taking $B_{n_1n_2}=1$ for a particular n_1, n_2 and zero for the rest, with the restriction that all the exponents are nonnegative. We note though that η_2 contains n_1 but not n_2 in its exponent. Thus there is the possibility that

$$\sum_{n_2}B'_{n_1n_2}(3,3)^{n_2}(2,2)^{(2\nu-L-3n_2)/2-2n_1} \tag{3.19}$$

for some coefficients of $B'_{n_1n_2}$ could be divisible by η_2 and in this case the exponent of η_2 can take negative values. To avoid this problem we note that the polynomial characterized by the epd (3, 0), i.e.,

$$(3,0)\equiv\sqrt{7}\sum_{mm'm''}(-)^{m''}\langle 22mm'|2-m''\rangle\eta_m\eta_{m'}\eta_{m''} \tag{3.20}$$

is related with those of (3.10) by

$$-(3\sqrt{3}/4)(3,3)^2=\eta_2^3(3,0)-\tfrac{3}{2}(2,2)(2,0)\eta_2^2+\tfrac{1}{4}(2,2)^3. \tag{3.21}$$

Thus if L is even, then n_2 is also even, i.e.,

$$n_2\equiv 2\mu \tag{3.22}$$

and we can express $(3,3)^{2\mu}$ in terms of $(3,0)^{\mu}$ and powers of the other epd's.

Therefore for $2\nu-L$ even, we can also write (3.18) as

$$P(\eta_m)=\sum_{n_1\mu}B''_{\mu n_1}\eta_2^{L-\nu+2n_1+3\mu}(2,2)^{(2\nu-L)/2-3\mu-2n_1}\times(3,0)^{\mu}(2,0)^{n_1}. \tag{3.23}$$

It seems at first sight that we have not avoided the problem of divisibility in (3.23) as we can write

$$2n_1+3\mu=\tau \tag{3.24}$$

and expressing n_1 in terms of τ we have now to ask whether

$$P(\eta_m)\equiv\sum_{\mu}B'''_{\mu\tau}(3,0)^{\mu}(2,0)^{(\tau-3\mu)/2} \tag{3.25}$$

is divisible by η_2 for some value of the coefficients $B'''_{\mu\tau}$. The answer to this last question is immediate. The polynomial (3.25) corresponds to $L=0$ as the two epd (3, 0), (2, 0) appearing in it have angular momentum zero. Thus if we can factorize it in the form

$$P(\eta_m)=\eta_2^{x}R(\eta_m), \tag{3.26}$$

where x is some positive integer, then

$$L_1R=0,\quad L_0R=-2xR. \tag{3.27}$$

This implies that R is a polynomial in the η_m's corresponding to a negative angular momentum which is clearly impossible.

We have then from (3.23) that the different polynomials are given by

$$P_{\nu\mu Ln_1}(\eta_m)\equiv\eta_2^{L-\nu+2n_1+3\mu}(2,2)^{(2\nu-L)/2-3\mu-2n_1}(3,0)^{\mu}(2,0)^{n_1}, \tag{3.28a}$$

when L is even and μ, n_1, L, ν are restricted by the fact that all exponents must be nonnegative.

A similar analysis for L odd, in which case n_2 is odd and can be written as

$$n_2=2\mu+1 \tag{3.29}$$

leads to

$$P_{\nu\mu L n_1}(\eta_m)$$
$$\equiv (3,3)\eta_2^{L-\nu+2n_1+3\mu}(2,2)^{(2\nu-L-3)/2-3\mu-2n_1}(3,0)^{\mu}(2,0)^{n_1}, \tag{3.28b}$$

where again μ, n_1, L, ν are restricted by the fact that all exponents must be nonnegative.

The polynomials (3.28), besides being characterized by the IR ν of U(5) and L of O(3), have two other labels, the nonnegative integers μ, n_1 which indicate the powers of the operators associated with the epd (2, 0), (3, 0).

We have obtained the complete, though not necessarily orthonormal, set of states of definite number of quanta ν and maximum projection of angular momentum $M=L$. For arbitrary M we just have to apply $(L_{-1})^{L-M}$ of (3.6) to the polynomials (3.28). The states (3.28) do not correspond though to a given seniority and in the next section we indicate how we can introduce this label in the classification scheme.

4. CONSTRUCTION OF STATES OF GIVEN SENIORITY THROUGH TRACELESS BOSON OPERATORS

The states belonging to the IR $(\lambda 0)$ of O(5) are eigenstates of Λ^2, the quadratic Casimir operator of the group, with eigenvalue $\lambda(\lambda+3)$,[13]

$$\Lambda^2 P(\eta_m)|0\rangle = \lambda(\lambda+3)P(\eta_m)|0\rangle. \tag{4.1}$$

The expression for Λ^2 is[10]

$$\Lambda^2 = \tfrac{1}{2}\sum_{mm'}[\eta_m\xi^{m'} - \eta^{m'}\xi_m][\eta_{m'}\xi^m - \eta^m\xi_{m'}], \tag{4.2}$$

and after some rearrangement of factors, it can be taken to the form

$$\Lambda^2 = N(N+3) - \left(\sum_{m'}\eta_{m'}\eta^{m'}\right)\left(\sum_m \xi_m\xi^m\right), \tag{4.3}$$

with N given in Eq. (3.5). From (4.3) we see that the eigenstates of Λ^2 with eigenvalue $\lambda(\lambda+3)$ have the form $P(\eta_m)|0\rangle$, where $P(\eta_m)$ is a homogeneous polynomial of degree λ in η_m, which is "harmonic", i.e.,

$$\sum_m \xi_m\xi^m P(\eta_\tau)|0\rangle = 0. \tag{4.4}$$

If we take the polynomials (3.28) with $\nu=\lambda$, we find that they do not satisfy the condition (4.4) as they stand. There is however a method, originated by Vilenkin[7] and further developed by Lohe,[6] by means of which we can enforce the condition (4.4) in a relatively simple way.

Following these authors we introduce "traceless boson operators" defined by

$$a_m^+ \equiv \eta_m - (2,0)(2N+5)^{-1}\xi_m, \quad m = 2,1,0,-1,-2, \tag{4.5}$$

where N is the number operator of Eq. (3.5), and (2, 0) is the second degree polynomial of angular momentum 0 of Eq. (3.10a), i.e., the one associated with two paired quanta. Using the identities

$$(2N+5)^{-1}\eta_m = \eta_m(2N+7)^{-1}, \tag{4.6a}$$

$$(2N+5)^{-1}\xi_m = \xi_m(2N+3)^{-1}, \tag{4.6b}$$

which hold when we apply the operators to homogeneous polynomials in the η's, we can easily show that

$$[a_m^+, a_{m'}^+] = 0, \tag{4.7}$$

and furthermore,

$$\sum_{m=2}^{-2} a_m^+ a^{+m} = (4N^2-1)^{-1}(2,0)^2 \sum_{m=2}^{-2}\xi_m\xi^m. \tag{4.8}$$

We now turn our attention to the polynomials $P_{\nu\mu L n_1}(\eta_m)$ of (3.28). If we replace η_m by a_m^+ in these polynomials and apply them to $|0\rangle$, we see from (4.8) that the corresponding states will vanish unless $n_1=0$. Assuming this last condition and taking $\nu=\lambda$, we have the states

$$|\lambda\mu L) \equiv P_{\lambda\mu L 0}(a_m^+)|0\rangle. \tag{4.9}$$

These states are linear combinations of terms like $\eta_{m_1}\eta_{m_2}\cdots\eta_{m_\lambda}|0\rangle$, i.e., they are homogeneous of degree λ in η_m, and they continue to be characterized by the angular momentum L.[6] Moreover, if we apply $\sum_m a_m^+ a^{+m}$ on the state $|\lambda\mu L)$ we obtain, from Eq. (4.7) and (4.8),

$$(4N^2-1)^{-1}(2,0)^2\sum_m \xi^m\xi^m|\lambda\mu L)$$
$$= \sum_m a_m^+ a^{+m} P_{\lambda\mu L 0}(a_{m'}^+)|0\rangle$$
$$= [P_{\lambda\mu L 0}(a_m^+)](4N^2-1)^{-1}(2,0)^2\sum_m \xi_m\xi^m|0\rangle = 0, \tag{4.10}$$

and since the factor $(4N^2-1)^{-1}(2,0)^2$ does not vanish identically, it follows that the states $P_{\lambda\mu L 0}(a^+)|0\rangle$ are "harmonic", i.e., they satisfy condition (4.4). Thus the states of Eq. (4.9) correspond to the IR $(\lambda 0)$ of O(5) and $[\lambda 0000]$ of U(5).

As the states (4.9) are eigenstates of H, Λ^2, L^2, and L_z with eigenvalues $\lambda+\frac{5}{2}$, $\lambda(\lambda+3)$, $L(L+1)$, and L, respectively, they obviously are identical to $|\lambda\mu L)$ of (2.25) if the extra label μ needed to characterize them denotes the number of triplets of traceless operators a_m^+ coupled to zero angular momentum. Introducing the notation $[\nu, L]$ for the elementary permissible diagrams (3.10), (3.20) where we replace η_m by a_m^+ as well as

$$[1,2] \equiv a_2^+, \tag{4.11}$$

we can express the states $|\lambda\mu L)$ in the operator form

$$|\lambda\mu L) = [1,2]^{L-\lambda+3\mu}[2,2]^{(2\lambda-L)/2-3\mu}[3,0]^{\mu}|0\rangle$$
$$\text{if } L \text{ is even}, \tag{4.12a}$$

$$|\lambda\mu L) = [1,2]^{L-\lambda+3\mu}[2,2]^{(2\lambda-3-L)/2-3\mu}[3,3][3,0]^{\mu}|0\rangle$$
$$\text{if } L \text{ is odd}, \tag{4.12b}$$

where all exponents must be nonnegative so the following inequalitites are satisfied:

$$L-\lambda+3\mu \geq 0, \quad \tfrac{1}{2}(2\lambda-L)-3\mu \geq 0, \quad \mu \geq 0$$
$$\text{for } L \text{ even}, \tag{4.13a}$$

$$L-\lambda+3\mu \geq 0, \quad \tfrac{1}{2}(2\lambda-L-3)-3\mu \geq 0, \quad \mu \geq 0$$
$$\text{for } L \text{ odd}, \tag{4.13b}$$

We show in Appendix A, that the inequalities (4.13) guarantee that the number of states (4.12) (of which there there are $2L+1$ for each L when we consider all possible values of M) for a given λ is

$$d_\lambda = \tfrac{1}{6}(\lambda+1)(\lambda+2)(2\lambda+3). \tag{4.14}$$

This is exactly the dimensionality d_λ of the IR $(\lambda 0)$ of O(5).[13] We note furthermore that the states (4.12) are linearly independent. This certainly is true for states of different L or λ as they are even orthogonal. For the same λ and L the polynomials in (4.12a), for which L is even, contain a_2^+ to the highest power $(L/2)+\mu$ as from (4.11), (3.10c), (3.22) we see that $[1,2]$, $[2,2]$, $[3,0]$ contain a_2^+ linearly. Thus no linear combination on the index μ of the polynomials (4.12a) can vanish and the same applies to those in (4.12b) for L odd.

The inequalities (4.13a) show that for $L=0$ we have $\lambda=3\mu$ and thus the state

$$|3\mu,\mu,0)=[3,0]^\mu\,|0\rangle \tag{4.15a}$$

must be identical to (2.26) if the constant A_μ in the latter is selected appropriately. Furthermore from (4.13b) we see that for $L=3$, $\lambda=3\mu+3$ and thus the state

$$|3\mu+3,\mu,3)=[3,3][3,0]^\mu\,|0\rangle \tag{4.15b}$$

must coincide with (2.31) for an appropriate selection of B_μ. Looking then back to the states $|\lambda\mu L)$ of (4.12) we see that all of them could be obtained from (4.15) if we apply powers of the operators $[1,2]$ and $[2,2]$.

The best way of getting $\phi_K^{\lambda\mu L}(\gamma)$ in the states $|\lambda\mu L)$ of (2.25) and (2.19) seems to be a recursive one starting from the state (2.26) of $L=0$ for L even, or (2.31) of $L=3$ for L odd. In Appendix B we applied the operator a_m^+ to the state $|\nu=\lambda,\lambda,\mu,L,M\rangle$ of (2.25) and found out that

$$\begin{aligned}a_m^+|\nu=\lambda,\lambda,\mu,L,M\rangle&=\left[\frac{2}{\Gamma(\lambda+\frac{7}{2})}\right]^{1/2}\beta^{\lambda+1}\exp(-\beta^2/2)\\&\times\sum_{\bar L\bar M\bar K}\left\{\sum_K\left[\langle L2Mm\,|\,\bar L\bar M\rangle\,Q_{K\bar K}^{\lambda L\bar L}\left(\gamma,\frac{\partial}{\partial\gamma}\right)\,\phi_K^{\lambda\mu L}(\gamma)\right]\right.\\&\left.\times[D^{\bar L*}_{\bar M\bar K}(\vartheta_i)+(-)^{\bar L}D^{\bar L*}_{\bar M-\bar K}(\vartheta_i)]\right\},\end{aligned} \tag{4.16}$$

where $Q^{\lambda L\bar L}_{K\bar K}(\gamma,\partial/\partial\gamma)$ is an operator function of γ and $\partial/\partial\gamma$ linear in the latter and given by (B.17).

Turning now our attention to the states (4.12) we note that

$$[1,2]\,|\lambda\mu L)=|\lambda+1,\mu,L+2), \tag{4.17a}$$

$$[2,2]\,|\lambda\mu L)=|\lambda+2,\mu,L+2). \tag{4.17b}$$

From the definitions (2.25), (2.19) of the states $|\lambda\mu L)$, and the explicit form of the operators $[1,2]$ and $[2,2]$, i.e.,

$$[1,2]=a_2^+ \tag{4.18a}$$

$$[2,2]=\sqrt{7}\sum_{mm'}\langle 22mm'\,|\,22\rangle\,a_m^+a_{m'}^+ \tag{4.18b}$$

we see from (4.16) and the elementary recoupling theory that[19]

$$\phi_{\bar K}^{\lambda+1,\mu,L+2}(\gamma)=\sum_K\,Q_{K\bar K}^{\lambda L,L+2}\left(\gamma,\frac{\partial}{\partial\gamma}\right)\,\phi_K^{\lambda\mu L}(\gamma), \tag{4.19a}$$

$$\begin{aligned}\phi_{K'}^{\lambda+2,\mu,L+2}(\gamma)=\sum_{\bar L\bar K K}&\left\{\sqrt{35(2\bar L+1)}\;W(L,L+2,2,2;2\bar L)\right.\\&\left.\times Q^{\lambda+1\bar L L+2}_{\bar K K'}\left(\gamma,\frac{\partial}{\partial\gamma}\right)\;Q^{\lambda L\bar L}_{K\bar K}\left(\gamma,\frac{\partial}{\partial\gamma}\right)\;\phi_K^{\lambda\mu L}(\gamma)\right\},\end{aligned} \tag{4.19b}$$

where W is a Racah coefficient. Thus we have given an explicit recursive procedure to get all states $|\nu\lambda\mu LM\rangle$ of (2.19) from those of $L=0$ if L is even, or $L=3$ if L is odd.

The states $|\nu\lambda\mu LM\rangle$ though complete are not orthogonal in the index μ and have not been normalized, but once obtained they can lead to an orthonormal set of states through an appropriate Hilbert–Schmidt procedure.

Having solved exactly the problem of the quadrupole vibrations of the liquid drop model of the nucleus, we briefly discuss in the concluding section the procedure for evaluating matrix elements as well as some possible applications.

5. CONCLUSION

The states $|\nu\lambda\mu LM\rangle$ of (2.19) with the $\phi_K^{\lambda\mu L}(\gamma)$ determined through (4.19), have many applications. To begin with we may consider extensions of the Lagrangian (2.5) in which besides the quadratic term in the potential

$$\{2,0\}\equiv\sum_m\,(-)^m\alpha_m\alpha_{-m}=\beta^2, \tag{5.1}$$

we have a cubic one of the form[9]

$$\begin{aligned}\{3,0\}\equiv\sqrt{7}\sum_{mm'm''}(-)^{m''}\langle 22mm'\,|\,2-m''\rangle\,\alpha_m\alpha_{m'}\alpha_{m''}\\=-\sqrt{2}\,\beta^3\cos3\gamma\end{aligned} \tag{5.2}$$

as well as higher power ones which are formed from products of $\{2,0\}$, $\{3,0\}$, i.e.,

$$\{2,0\}^2;\{2,0\}\,\{3,0\};\{2,0\}^3;\{3,0\}^2,\ \text{etc.} \tag{5.3}$$

This type of Hamiltonian has been applied to transitional nuclei[20] and considered also in relation to higher order deformations.

The simplest way of dealing with Hamiltonians containing terms of the form (5.3) would be to calculate their matrix elements with respect to the states $|\nu\lambda\mu LM\rangle$ and diagonalize the corresponding matrix for which L^2, L_z remain good integrals of motion. The matrix elements of powers of β with respect to the $F_n^\lambda(\beta)$ of (2.20) are well known.[14] Those of powers of $\cos3\gamma$ with respect to the $\phi_K^{\lambda\mu L}(\gamma)$ can be determined from the explicit forms of the latter functions given by the recursion relations, (4.19).

A similar problem concerns the determination of the transition probabilities between states of the type (2.19), when we have a multipole operator such as α_m or appropriate powers of it.[9] Again it is a question of determining matrix elements, though now we also have the Wigner functions $D^L_{mm'}(\vartheta_i)$ involved in the operator α_m as seen from (2.3).

Recently states of n bosons, of the type η_m in (3.2), characterized by the IR of the chain of groups $U(5)\supset O(5)\supset O(3)$, have been considered in the analysis of elementary excitations in vibrational nuclei by Arima and Iachello.[21] All of the discussion in Sec. 3,4 of this paper applies also to them. In particular the states required in Ref. 21 can be written in the form

$$|\nu\lambda\mu L, M=L\rangle=(2,0)^{(\nu-\lambda)/2}|\lambda\mu L), \tag{5.4}$$

where the epd (2, 0) is given by (3.10a) and the kets $|\lambda\mu L)$, have the operator form (4.12).

Finally we note that while the analysis carried out in the present paper concerns only quadrupole vibrations, i.e., $l=2$, it is in principle generalizable to other l. For example for octupole vibrations $l=3$, we have the chain of groups $U(7)\supset O(7)\supset O(3)$. The determination of the polynomials in the η_m's, where now $m=3,2,1,0,-1,-2,-3$, in the chain $U(7)\supset O(3)$ can be done along lines similar to those in Sec. 3 for $U(5)\supset O(3)$, but the divisibility problem is likely to be much more difficult. The introduction of an appropriate traceless boson operator will then permit the characterization of the states by the IR of O(7), in a way is similar to what was done in Sec. 4 for O(5).

Applications and extensions of the present developments will be published elsewhere.

ACKNOWLEDGMENT

The authors would like to dedicate this paper to Professor V. Bargmann on the occasion of his retirement from Princeton University. Professor Bargmann's advice on problems of mathematical physics has been so freely given that the recipients frequently forget the origin of their ideas. No doubt he will recognize outgrowths of his own in this and many other papers.

APPENDIX A

We shall give here the proof that the sum of the dimensions of all the IR of O(3) allowed by the inequalities in Eqs. (4.13), for a fixed λ, is equal to the dimension of the IR $(\lambda 0)$ of O(5).

The inequalities (4.13a) can be collected in

$$0\leq\lambda-3\mu\leq L\leq 2\lambda-6\mu,\quad L \text{ even}, \tag{A1a}$$

while those in Eq. (4.13b) give

$$0\leq\lambda-3\mu\leq L\leq 2\lambda-6\mu-3,\quad L \text{ odd} \tag{A1b}$$

with μ being a nonnegative integer in both cases. From here it follows that the sum we must evaluate is

$$S=\sum_{\mu=0}^{[\lambda/3]}\sum_{L=2[(\lambda+1-3\mu)/2]}^{2\lambda-6\mu}{}'(2L+1) + \sum_{\mu=0}^{[(\lambda-3)/3]}\sum_{L=2[(\lambda-3\mu)/2]+1}^{2\lambda-6\mu-3}{}'(2L+1), \tag{A2}$$

where $[x]$ is the largest integer contained in x, and a prime means that the sum goes by steps of two.

We can regroup terms and write the sum (A2) as

$$S=\sum_{\mu=0}^{[(\lambda-3)/3]}\left[\sum_{L=\lambda-3\mu}^{2\lambda-6\mu-2}(2L+1)+2(2\lambda-6\mu)+1\right] + \begin{Bmatrix}1\\5\\14\end{Bmatrix}, \tag{A3}$$

where the last bracket is the contribution of the terms with $\mu=[\lambda/3]$ and contains, respectively, the dimensions of the following IR of O(3):

$$\begin{aligned} &L=0 \quad \text{when} \quad \lambda\equiv 0 \mod 3,\\ &L=2 \quad \text{when} \quad \lambda\equiv 1 \mod 3,\\ &L=2,4 \quad \text{when} \quad \lambda\equiv 2 \mod 3. \end{aligned} \tag{A4}$$

Using the auxiliary formula,

$$\sum_{L=\alpha}^{\beta}(2L+1)=(\beta+1)^2-\alpha^2, \tag{A5}$$

we have

$$S=\sum_{\mu=0}^{[(\lambda-3)/3]}(3\lambda^2+2-18\lambda\mu+27\mu^2)+\begin{Bmatrix}1\\5\\14\end{Bmatrix}. \tag{A6}$$

The upper limit of the sum for the three cases of (A4) is, respectively, $(\lambda-3)/3$, $(\lambda-4)/3$, $(\lambda-5)/3$. The sum can be effected in each case using the formulas $\sum_{n=0}^{k}n=\frac{1}{2}k(k+1)$ and $\sum_{n=0}^{k}n^2=\frac{1}{6}k(k+1)(2k+1)$, giving in all three cases

$$S=\tfrac{1}{6}(\lambda+1)(\lambda+2)(2\lambda+3), \tag{A7}$$

which agrees with the dimension of the IR$(\lambda 0)$ of O(5).

From (A3) and (A4) it can be deduced that, whatever the value of λ, $L=1$ never occurs, and $L=0,2,3,4,5,7$ occurs at most once.

APPENDIX B

In this appendix we shall give an outline of the derivation of the operator $Q^{\lambda L\bar{L}}_{K\bar{K}}(\gamma,\partial/\partial\gamma)$ of Eq. (4.16).

The traceless boson operator a^+_m was defined in Eq. (4.5) as

$$a^+_m=\eta_m-(2,0)(2N+5)^{-1}\xi_m. \tag{B1}$$

It is, however, convenient to use the commutation relations of the η's and ξ's to move the ξ_m in (B1) to the left and obtain the following expression, equivalent to (B1) when applied on a homogeneous polynomial in η_m,

$$a^+_m=\eta_m\left(\frac{2N+5}{2N+3}\right)-\xi_m(2,0)(2N+3)^{-1}. \tag{B2}$$

Introducing here the relations

$$\eta_m=(1/\sqrt{2})(\alpha_m-i\pi_m),\quad \xi_m=(1/\sqrt{2})(\alpha_m+i\pi_m), \tag{B3}$$

with α_m, π_m given in Eqs. (2.8), and (2.9), and furthermore, taking into account that a^+_m is going to be applied on a state (2.19) which we write as

$$|\nu=\lambda,\lambda\mu LM\rangle=\left[\frac{2}{\Gamma(\lambda+\frac{5}{2})}\right]^{1/2}\beta^\lambda\exp(-\beta^2/2)\,\chi(\gamma,\vartheta_i), \tag{B4}$$

we conclude that

$$a^+_m|\nu=\lambda,\lambda\mu LM\rangle =\left[\frac{2}{\Gamma(\lambda+\frac{7}{2})}\right]^{1/2}\beta^{\lambda+1} \times\exp(-\beta^2/2)\,\frac{(2\lambda+5)^{1/2}}{(2\lambda+3)}\left[(\lambda+3)\frac{\alpha_m}{\beta}-i\Delta_m\right]\chi(\gamma,\vartheta_i) \tag{B5}$$

where we made use of the fact that

$$(2,0)=\tfrac{1}{2}\sum_{m=2}^{-2}(\alpha_m\alpha^m-\pi_m\pi^m-i\pi^m\alpha_m-i\alpha_m\pi^m)$$

$$=\beta^2-N-5-\beta\frac{\partial}{\partial\beta}. \tag{B6}$$

For computational convenience we shall split the operator Δ_m of Eq. (2.9b) into two terms, namely

$$\Delta_m=V_m+W_m, \tag{B7a}$$

where

$$V_m=i\sum_{k\tau\rho}\langle 2\rho|L_k|2\tau\rangle^* D^{2*}_{m\rho}(\vartheta_i)\,I_k^{-1}L'_k\,a_\tau/\beta, \tag{B7b}$$

$$W_m=-(i/\sqrt{2})[D^{2*}_{m2}+D^{2*}_{m-2}]\cos\gamma\frac{\partial}{\partial\gamma}+iD^{2*}_{m0}\sin\gamma\frac{\partial}{\partial\gamma}. \tag{B7c}$$

As the effect of α_m and W_m on $\chi(\gamma,\vartheta_i)$ can be easily deduced, we shall concentrate our attention on the application of V_m on $\chi(\gamma,\vartheta_i)$.

Writing χ as

$$\chi(\gamma,\vartheta_i)=\chi_+ +(-)^L\chi_-,\quad \chi_\pm=\sum_K\phi_K^{\lambda\mu L}(\gamma)\,D^{L*}_{M\pm K}(\vartheta_i), \tag{B8}$$

we have, owing to the completeness of the set D^L_{MK},

$$V_m\chi_\pm=\sum_{\bar{L}\bar{M}\bar{K}}A_\pm^{\bar{L}\bar{M}\bar{K}}D^{\bar{L}*}_{\bar{M}\bar{K}}, \tag{B9}$$

with

$$A_\pm^{\bar{L}\bar{M}\bar{K}}=[(2\bar{L}+1)/8\pi^2]\int D^{\bar{L}}_{\bar{M}\bar{K}}(\vartheta_i)\,V_m\chi_\pm\,d\Omega$$

$$=i\langle L2Mm|\bar{L}\bar{M}\rangle\sum_{K\bar{L}\tau}\left[\sum_{k\rho K'}\langle 2\rho|L_k|2\tau\rangle^*\times\langle LK'|L_k|L,\pm K\rangle^*\langle L2K'\rho|\bar{L}\bar{K}\rangle I_k^{-1}\right]\left(\frac{a_\tau}{\beta}\right)\phi_K^{\lambda\mu L}(\gamma), \tag{B10}$$

where in the last step use was made of Eq. (4.62) of Ref. 19 as well as the fact that

$$L'_kD^{L*}_{M\pm K}=\sum_{K'}\langle LK'|L_k|L,\pm K\rangle^* D^{L*}_{MK'}. \tag{B11}$$

We note now that the term in the square bracket of (B10) could be written as

$$\sum_{L''K''}\langle L2\bar{L}\bar{K}|\sum_k I_k^{-1}L_k^{(1)}L_k^{(2)}|L2L''K''\rangle\langle L2,\pm K\tau|L''K''\rangle \tag{B12}$$

where we distinguish the two operators L_k appearing in (B10) by the indices 1 and 2. We furthermore have the identity

$$\sum_{k=1}^{3}I_k^{-1}L_k^{(1)}L_k^{(2)}=\tfrac{1}{2}(I_1^{-1}-I_2^{-1})\{[\mathbf{L}^{(1)}\times\mathbf{L}^{(2)}]^2_2+[\mathbf{L}^{(1)}\times\mathbf{L}^{(2)}]^2_{-2}\}$$

$$+\tfrac{1}{2}(2I_3^{-1}-I_2^{-1}-I_1^{-1})\sqrt{2/3}\,[\mathbf{L}^{(1)}\times\mathbf{L}^{(2)}]^2_0$$

$$-(3\sqrt{3}/4)\frac{1}{\sin^2 3\gamma}\,[\mathbf{L}^{(1)}\times\mathbf{L}^{(2)}]^0_0, \tag{B13}$$

where

$$[\mathbf{L}^{(1)}\times\mathbf{L}^{(2)}]^j_m\equiv\sum_{m'm''}\langle 11m'm''|jm\rangle L^{(1)}_{m'}L^{(2)}_{m''}. \tag{B14}$$

Introducing (B13) in (B12) and using standard recoupling techniques, we finally write

$$A_\pm^{\bar{L}\bar{M}\bar{K}}=i\langle L2Mm|\bar{L}\bar{M}\rangle\sum_{L''K''}\{\langle L2\bar{L}\bar{K}|\sum_k I_k^{-1}L_k^{(1)}L_k^{(2)}|L2L''K''\rangle\times\sum_{K\tau}\langle L2\pm K\tau|L''K''\rangle\left(\frac{a_\tau}{\beta}\right)\phi_K^{\lambda\mu L}(\gamma)\} \tag{B15a}$$

with

$$\langle L2\bar{L}\bar{K}|\sum_k I_k^{-1}L_k^{(1)}L_k^{(2)}|L2L''K''\rangle$$

$$=\frac{5[6L(L+1)(2L+1)(2L''+1)]^{1/2}}{\sin^2 3\gamma}\begin{Bmatrix}L2\bar{L}\\L2L''\\112\end{Bmatrix}$$

$$\times[-2\sqrt{3}\sin^3\gamma\cos\gamma(\langle L''2K''2|\bar{L}\bar{K}\rangle+\langle L''2K''-2|\bar{L}\bar{K}\rangle)$$

$$+(1/\sqrt{6})(3[\tfrac{1}{2}+\cos 2\gamma]^2-\tfrac{9}{4})\langle L''2K''0|\bar{L}\bar{K}\rangle]$$

$$-\frac{3}{8\sin^2 3\gamma}[L(L+1)+6-\bar{L}(\bar{L}+1)]\delta_{\bar{L}L''}\delta_{\bar{K}K''} \tag{B15b}$$

where $\{\}$ is a $9j$ coefficient.

Carrying out a similar analysis for α_m/β and W_m we see that

$$a_m^\dagger|\nu=\lambda,\lambda\mu LM\rangle$$

$$=\left[\frac{2}{\Gamma(\lambda+\frac{5}{2})}\right]^{1/2}\beta^{\lambda+1}\exp(-\beta^2/2)\sum_{\bar{L}\bar{M}\bar{K}}\langle L2Mm|\bar{L}\bar{M}\rangle\times\sum_K\left[Q^{\lambda L\bar{L}}_{K\bar{K}}\left(\gamma,\frac{\partial}{\partial\gamma}\right)\phi_K^{\lambda\mu L}(\gamma)\right]2D^{\bar{L}*}_{\bar{M}\bar{K}}(\vartheta_i), \tag{B16}$$

where

$$Q^{\lambda L\bar{L}}_{K\bar{K}}\left(\gamma,\frac{\partial}{\partial\gamma}\right)=\frac{(2\lambda+5)^{1/2}}{2(2\lambda+3)}$$

$$\times\Bigg\{\sum_\tau\Bigg[\sum_{L''K''}(\langle L2\bar{L}\bar{K}|\sum_k I_k^{-1}L_k^{(1)}L_k^{(2)}|L2L''K''\rangle$$

$$+(\lambda+3)\,\delta_{L''\bar{L}}\,\delta_{K''\bar{K}})(\langle L2K\tau|L''K''\rangle$$

$$+(-)^L\langle L2-K\tau|L''K''\rangle)\left(\frac{a_\tau}{\beta}\right)\Bigg]-(1/\sqrt{2})(\langle L2K2|\bar{L}\bar{K}\rangle$$

$$+(-)^L\langle L2-K-2|\bar{L}\bar{K}\rangle+\langle L2K-2|\bar{L}\bar{K}\rangle$$

$$+(-)^L\langle L2-K2|\bar{L}\bar{K}\rangle)\cos\gamma\frac{\partial}{\partial\gamma}+(\langle L2K0|\bar{L}\bar{K}\rangle$$

$$+(-)^L\langle L2-K0|\bar{L}\bar{K}\rangle)\sin\gamma\frac{\partial}{\partial\gamma}\Bigg\}, \tag{B17}$$

and $\langle\cdots|\sum_k I_k^{-1}L_k^{(1)}L_k^{(2)}|\cdots\rangle$ is given in Eq. (B15b).

We note now that

$$Q^{\lambda L\bar{L}}_{K\bar{K}}=(-)^{\bar{L}}\,Q^{\lambda L\bar{L}}_{K-\bar{K}}, \tag{B18}$$

as can be seen from (B17) changing the sign of all repeated magnetic quantum numbers of the Clebsch–Gordan coefficients and using their symmetry properties. Thus we conclude that in the expansion (B16) we could replace

$$2D^{\bar{L}*}_{\bar{M}\bar{K}}(\vartheta_i)\to D^{\bar{L}*}_{\bar{M}\bar{K}}(\vartheta_i)+(-)^{\bar{L}}D^{\bar{L}*}_{\bar{M}-\bar{K}}(\vartheta_i),$$

and the operator (B17) is the one we require in Eq. (4.16). We note incidentally, from the Wigner coefficients in (B17) and the fact that $\tau=\pm 2,0$, that all $\bar{K}$ will be even if K is even, thus keeping the restriction mentioned in (2.21).

Note added in proof: It has been brought to our attention that particular solutions of the problem of quadrupole vibrations of the nucleus have been obtained for other L besides 0 and 3. D. Bes[22] considered the set of coupled differential equations for $\phi_K^{\lambda\mu L}(\gamma)$ when $L=2,4,5,6$. He gave explicit solutions for small values

of λ in terms of polynomials of $\cos 3\gamma$. Budnik, Rabotnov, and Seregin[23] gave for all λ the explicit solution when $L=2$.

*Member of the Instituto Nacional de Energía Nuclear.
†Member of El Colegio Nacional.
‡On leave of absence from the University of McGill, Montreal, Canada.

[1]A. Bohr, Kgl. Dan. Videnskab. Selsk. Mat. Fys. Medd. **26**, 14 (1952).
[2]A. Bohr and B. Mottelson, Kgl. Dan. Videnskab. Selsk. Mat. Fys. Medd. **27**, 16 (1953).
[3]A. Bohr, "Rotational States in Atomic Nuclei," Thesis, Copenhagen (1954).
[4]A.S. Davidov, Nucl. Phys. **24**, 682 (1961).
[5]V. Bargmann and M. Moshinsky, Nucl. Phys. **23**, 177 (1961).
[6]M.A. Lohe, "The Development of the Boson Calculus for the Orthogonal and Symplectic Groups," Thesis, University of Adelaide (1974).
[7]N.Y. Vilenkin, *Special Functions and Theory of Groups Representations* (A.M.S. Transl., Providence, R.I., 1968).
[8]G. Racah, Phys. Rev. **63**, 367 (1943).
[9]J.M. Eisenberg and W. Greiner, *Nuclear Models* (North-Holland, Amsterdam, 1970), Chap. 2, pp. 35—36.
[10]Ref. 9, Appendix A, p. 431.
[11]Ref. 9, Chap. 5, pp. 103, 96, 106.
[12]Ref. 9, Chap. 6, pp. 130, 131, 149.
[13]J.D. Louck and H.W. Galbraith, Rev. Mod. Phys. **44**, 540 (1972).
[14]M. Moshinsky, T.H. Seligman, and B. Wolf, J. Math. Phys. **13**, 901 (1972).
[15]F. Calogero, J. Math. Phys. **12**, 2191 (1969).
[16]M. Moshinsky, *Group Theory and the Many Body Problem* (Gordon and Breach, New York, 1967).
[17]M. Moshinsky and V. Syamala Devi, J. Math. Phys. **10**, 455 (1969).
[18]R.T. Sharp and C.S. Lam, J. Math. Phys. **10**, 2033 (1969).
[19]M.E. Rose, *Elementary Theory of Angular Momentum* (Wiley, New York, 1957).
[20]L. von Bernus *et al.* "A Collective Model for Transitional Nuclei," in *Heavy Ion, High Spin States and Nuclear Structure* (International Atomic Energy Agency, Vienna, 1975).
[21]A. Arima and F. Iachello, Phys. Lett. B **57**, 309 (1974); Phys. Lett. B **57**, 39 (1975).
[22]D.R. Bès, Nucl. Phys. **10**, 373 (1959).
[23]A.P. Budnik, N.S. Rabotnov, and A.A. Seregin, Yad. Fiz. 477 (1970) [Sov. J. Nucl. Phys. **12**, 261 (1971)].

An approximant for democratic representation of all Born terms

Robert C. Brunet

Département de Mathématiques, Université de Montréal, Montréal, Québec, Canada
(Received 11 September 1975; revised manuscript received 21 October 1975)

We propose an approximant which attempts to reconstruct a solution starting from the Born terms of a formal power series. The approximant follows closely the Fredholm solutions to Born–Neumann series of completely continuous operators and their finite rank approximations. We show that if the Fredholm solution is written as a combination of Born terms, it tends to become independent of the expansion parameter λ asymptotically. The proposed approximants then appear as solutions to differential equations approximately satisfied by the formal power series, a feature they share with kernel of finite rank approximations. All Born terms are put on equal footing and their respective weight is determined independently of λ by initial conditions; thus knowledge of the solution and $(N-1)$ derivatives at one point is necessary. Analytic and crossing-symmetry properties are preserved by the proposed approximant, but unitarity is not insured and has to be examined specifically. Its error structure and its properties are studied and compared to Padé approximants.

1. INTRODUCTION

One of the outstanding practical problems met in physics is the reconstruction of the true solution $f(x,\lambda)$ from the more or less formal power series known to physicists as a Born series and to mathematicians as a Neumann series:

$$\Sigma_f(x,\lambda)=f_0(x)+\lambda f_1(x)+\lambda^2 f_2(x)+\lambda^3 f_3(x)+\cdots . \tag{1.1}$$

This formal power series is usually all that is available because most theoretical schemes are only capable of yielding the "perturbative" Born terms through some iteration procedure.

The presently popular approach, when λ is known and relatively large, is the Padé approximant which concentrates on the λ dependence of $f(x,\lambda)$ and sees the Born terms $f_i(x)$ as coefficients of the power expansion in λ. Padé approximants have been successful in many instances. The literature on the subject is extensive, and there are recent review books.[1–3] The "x" dependence (possibly more than one variable, e.g., energy, momentum transfer, etc.) of the Born terms usually appears, however, in the denominator of the Padé approximant. This often introduces unwanted singularities in the "x" variable. Further the crossing properties of amplitudes, an important feature in high energy physics, are lost in the Padé approximant.

Our purpose here will be to tackle the reconstruction of $f(x,\lambda)$ starting also from the Born–Neumann series (1.1) but stressing instead the hitherto neglected "x" dependence. The approximant we wish to propose will seek to use to the maximum the information contained in the functional form in "x" of every Born term available. This should be particularly useful when the expansion parameter λ is unknown or when it is so large as to make it a purely formal device in obtaining a power series. We will define, from the Born terms, systems of differential equations, the solutions of which will constitute our approximants. But, whoever says differential equations implies initial conditions or boundary values; knowledge of these will replace that of the expansion parameter λ which we feign to ignore.

The perturbative and Padé methods have established a hierarchy among Born terms associated with the corresponding power in λ. To establish "democracy" among Born terms and discredit the importance of λ, especially when it is large, we will examine Born–Neumann series generated by a completely continuous integral operator and their finite rank approximations. In Sec. 2 we will study the Fredholm solutions of the latter. Those solutions are linear combinations of Born terms, and it turns out that the coefficients of the Born terms tend to become independent of λ asymptotically. In Sec. 3 we formally define our approximant using Wronskians and initial conditions to determine the weight of each Born term. To this effect we need to know the solution and $(N-1)$ of its derivatives at one point. In Sec. 4 we study its error structure, and compare some of its advantages and drawbacks with respect to the Padé approximant.

2. ON THE COMPLETELY CONTINUOUS INTEGRAL OPERATOR

Let us take an integral equation with a complete continuous operator K of kernel $k(x,x')$ on a Hilbert space:

$$(I-\lambda \mathrm{K})f=g. \tag{2.1}$$

This equation can be solved at least formally as a Born–Neumann series:

$$f(x,\lambda)-g(x)=\lambda f_1(x)+\lambda^2 f_2(x)+\lambda^3 f_3(x)\cdots . \tag{2.2}$$

The "perturbative" terms are obtained by repeated application of the operator K, thus:

$$f_j(x)=\mathrm{K}^j g. \tag{2.3}$$

This expansion by itself is of little use when $\|\lambda \mathrm{K}\|>1$; nonetheless, these perturbative or Born terms are often the only building blocks available in many physical problems.

As a first step we will relate these Born terms to the Fredholm resolvent of the equation. Following closely the notation of Byron and Fuller,[4] we have

$$f(x,\lambda)-g(x)=[\lambda/D(\lambda)]\int N(x,x',\lambda)g(x')\,dx'. \tag{2.4}$$

Both $N(x,x',\lambda)$ and $D(\lambda)$ are infinite series which converge for all λ. The Fredholm determinant is

$$D(\lambda)=\sum_{k=0}^{\infty}\lambda^k\frac{(-1)^k}{k!}d_k, \tag{2.5}$$

where d_k are constants (i.e., $d_0=1$, etc.). Using the notation $\mathrm{N}(\lambda)$ for the operator $\int dk' N(x,x',\lambda)$, we write

$$\mathrm{N}(\lambda)=\sum_{n=0}^{\infty}\lambda^n\frac{(-1)^n}{n!}\mathrm{N}_n. \tag{2.6}$$

Formally the operator $\mathrm{N}(\lambda)$ satisfies the equation

$$\mathrm{N}(\lambda)=D(\lambda)\mathrm{K}[I-\lambda\mathrm{K}]^{-1}. \tag{2.7}$$

Expanding $D(\lambda)$ and $[I-\lambda\mathrm{K}]^{-1}$ in powers of λ, and regrouping these powers, the expansion operators N_n of (2.6) can be written as

$$\mathrm{N}_n=\sum_{j=0}^{n}(-1)^j\frac{n!}{(n-j)!}d_{n-j}\mathrm{K}^{j+1}. \tag{2.8}$$

We can then use the Fredholm solution (2.4), again as the ratio of series converging for all λ, but this time using explicitly the Born terms of the formal power series (2.2),

$$f(x,\lambda)-g(x)=\frac{\lambda}{D(\lambda)}\sum_{n=0}^{\infty}\lambda^n\left(\sum_{j=0}^{n}\frac{(-1)^{n-j}}{(n-j)!}d_{n-j}f_{j+1}(x)\right). \tag{2.9}$$

At this point we replace the general kernel $k(x,x')$ by the approximating kernel of finite rank "L":

$$k^L(x,x')=\sum_{i=1}^{L}\alpha_i(x)\beta_i(x'). \tag{2.10}$$

(The superscript label "L" will be used to identify all objects pertaining to a finite rank kernel like the above.) The determinantal integrands entering in the definition of the classical Fredholm terms d_k and N_n will insure that $d_k^L\equiv 0$ for $k>L$ and $\mathrm{N}_n^L\equiv 0$ for $n>L-1$. The numerator and denominator series of (2.9) will then terminate, and we have

$$f^L(x,\lambda)-g(x)=\frac{\lambda}{D^L(\lambda)}\sum_{n=0}^{L-1}\lambda^n\left(\sum_{j=0}^{n}\frac{(-1)^{n-j}}{(n-j)!}d_{n-j}^L f_{j+1}^L(x)\right). \tag{2.11}$$

After rearranging the sums and regrouping the coefficients of the Born terms, we obtain

$$f^L(x,\lambda)-g(x)=\sum_{i=1}^{L}c_i^L(\lambda)f_i^L(x), \tag{2.12}$$

where

$$c_i^L(\lambda)=\frac{\lambda^i}{D^L(\lambda)}\sum_{l=0}^{L-i}\frac{(-1)^l}{l!}\lambda^l d_l^L, \tag{2.13}$$

and

$$D^L(\lambda)=\sum_{k=0}^{L}\frac{(-1)^k}{k!}\lambda^k d_k^L. \tag{2.14}$$

It will be no surprise, of course, that the "x" dependence of the solution $f^L(x,\lambda)$ is described by a linear combination of L Born terms for a kernel of finite rank L. Indeed each $f_i^L(x)$ is made up of a linear combination of the L independent functions $\alpha_i(x)$ of (2.10) through repeated application of K^L. Thus in general the solution $[f^L(x,\lambda)-g(x)]$ which is a linear combination of the L functions $\alpha_i(x)$ can equally be given as a linear combination of L Born terms $f_i^L(x)$. Should some fortuitous linear dependence occur between the first L terms $f_i^L(x)$, the adjunction of further $f_i^L(x)$ should restore the generality of the solution.

This could have all been said without working through to equations (2.13) and (2.14), but we derived them in order to study the λ dependence of the coefficients $c_i^L(\lambda)$ of the Born terms in (2.12). When λ is small, the coefficients $c_i^L(\lambda)$ obviously have a leading power λ^i corresponding to their index "i", and

$$f^L(x,\lambda)=\lambda[1+O(\lambda)]f_1^L(x)+\lambda^2[1+O(\lambda)]f_2^L(x)+\cdots+\lambda^L[1+O(\lambda)]f_L^L(x). \tag{2.15}$$

This reestablishes as expected the hiearchy of importance among the Born terms, and then the weight to be granted to each $f_i^L(x)$ is almost equivalent to the corresponding power in λ.

However, we wish to stress that when λ is large, the situation is entirely different. In fact all the coefficients $c_i^L(\lambda)$ tend asymptotically to become independent of λ, and most importantly this is true for any value of L one might have chosen to approximate the original kernel K:

$$\lim_{\lambda\to\infty}c_i^L(\lambda)\to(-1)^i[L!/(L-i)!]\,d_{L-i}^L/d_L^L. \tag{2.16}$$

Hence when λ is large, the hiearchy among Born terms is abolished since the weights to be attributed to each $f_i^L(x)$ are roughly independent of λ and of the same order of magnitude. The weights $c_i^L(\lambda)$ are in fact complicated functions of λ which may be unravelled only if we have L further Born terms on hand. That this can be done through a $[L/L]$ Padé approximant when $2L$ Born terms are available was shown by Chisolm,[5] and then the problem is completely solved in both its "x" and λ dependence. It is important to note, however, that only with $2L$ Born terms in a $[L/L]$ Padé, and solely for a kernel of finite rank L, will the subtle cancellations of the "x" dependence take place in the denominator of the Padé approximant. If these conditions are not all met, stray "x" dependence in the Padé denominator will occur in contradiction to the Fredholm denominator (2.14). The case $L=1$ will illustrate this point. Given

$$k^I(x,x')=\alpha(x)\beta(x') \tag{2.17}$$

and the corresponding formal power series

$$f^I(x,\lambda)-g(x)=\lambda\alpha(x)(\beta,g)+\lambda^2\alpha(x)(\beta,\alpha)(\beta,g)+\cdots+\lambda^n\alpha(x)(\beta,\alpha)^{n-1}(\beta,g)\cdots, \tag{2.18}$$

where (β,α) and (β,g) are scalar products.

With two Born terms a Padé approximant $[1/1]$ gives the exact answer:

$$[1/1]_{f^I}-g(x)=\frac{\lambda f_1^I(x)}{1-\lambda f_2^I(x)/f_1^I(x)}=\frac{\lambda}{1-\lambda(\beta,\alpha)}f_1^I(x). \tag{2.19}$$

Note that the ratio $f_2^I(x)/f_1^I(x)$ just cancels the $\alpha(x)$ dependence in this simple case and serves to provide the constant term (β,α), but in general such cancellation of x dependence in the denominator does not take place. It is also interesting to note that consecutive Born terms $f_{n+1}^I(x)/f_n^I(x)$ would do. In view of this we would rather

take the outlook that when λ is large the coefficient is practically independent of λ, and that the correct "x" dependence will be given with a single Born term (anyone of them in fact):

$$f^I(x,\lambda)-g(x)=c_1^I f_1^I(x). \tag{2.20}$$

Let us now return to an idealized finite rank operator (which we need never know) and suppose it faithful enough to the original $k(x,x')$ so that each of the $f_i^L(x)$ of (2.12) will be fairly close to the actually available Born terms $f_i(x)$ of (2.2). Thus

$$f^L(x,\lambda)-g(x)\approx\sum_{i=1}^{L}c_i^L(\lambda)f_i(x). \tag{2.21}$$

Since our aim is not to approximate $f^L(x,\lambda)$ anyway but rather $f(x,\lambda)$, we may be bold enough to suppose that the substitution $f_i^L(x,\lambda)\to f_i(x,\lambda)$ goes in the right direction. This heuristic line of reasoning leads us to propose the following approximant to $f(x,\lambda)$:

$$A^N(x)-g(x)=\sum_{i}^{N}a_i^N f_i(x), \tag{2.22}$$

made out of N available Born terms.

This linear combination allows us (i) to bypass the difficult problem of the λ dependence of the coefficients which is less and less relevant as λ gets larger, (ii) to avoid stray "x" dependence in the denominator. This linear combination is surely reasonable whenever the generating kernel is well approximated by a kernel of finite rank. On the other hand, the coefficients of that linear combination will have to be determined in some empirical way. This really comes down to management of the available information; since Born terms are usually very limited in numbers, we are trying to use them in a more economical manner.

3. THE APPROXIMANT

Consideration of the completely continuous operator and its finite rank approximations led us to suggest an approximant $A^N(x)$ to $f(x,\lambda)$ where the N Born terms, available from a formal power series (2.2), are *a priori* on equal footing:

$$\tilde{A}^N(x)\equiv A^N(x)-g(x)=\sum_{i=1}^{N}a_i^N f_i(x). \tag{3.1}$$

Before we discuss the problem of determining the a_i^N, there is still a problem lingering: How many Born terms are sufficient to make a reasonable approximation of the "x" dependence? The criteria will have to come from the Born terms themselves since they are the only fully known objects we have. Our test will yield differential systems whose solutions will be of the type (3.1), but at the same time it will indicate how many Born terms are needed.

Given a Born–Neumann series where for convenience the λ dependence has been deleted in f and $\tilde{f}$,

$$\tilde{f}(x)\equiv f(x)-g(x)=\lambda f_1(x)+\lambda^2 f_2(x)+\cdots+\lambda^n f_n(x)\cdots, \tag{3.2}$$

we first divide by one of the Born terms, let us say $f_1(x)$:

$$\tilde{f}(x)/f_1(x)=\lambda+\lambda^2[f_2(x)/f_1(x)]+\lambda^3[f_3(x)/f_1(x)+\cdots+\lambda^n[f_n(x)/f_1(x)]. \tag{3.3}$$

If we were dealing with an operator of finite rank $L=1$, then the right-hand side of (3.3) would necessarily be a constant. Thus, if we were to find these ratios of Born terms nearly constant, a reasonable ansatz for an approximant would be

$$A^{N=1}(x)=a_1^{N=1}f_1(x). \tag{3.4}$$

If these ratios of Born terms are not constant, we first differentiate everywhere and then divide by one of the new coefficients of the powers of λ so obtained, say $[f_2(x)/f_1(x)]'$:

$$\frac{(\tilde{f}/f_1)'}{(f_2/f_1)'}=\lambda^2+\lambda^3\frac{W_2(f_1,f_3)}{W_2(f_1,f_2)}+\lambda^4\frac{W_2(f_1,f_4)}{W_2(f_1,f_2)}+\cdots+\lambda^n\frac{W_2(f_1,f_n)}{W_2(f_1,f_2)}\cdots. \tag{3.5}$$

However, the derivatives of the ratios of Wronskians on the rhs are such that

$$\left[\frac{W_2(f_1,f_n)}{W_2(f_1,f_2)}\right]'=\frac{W_2[W_2(f_1,f_2),W_2(f_1,f_n)]}{[W_2(f_1,f_2)]^2}=\frac{f_1W_3(f_1,f_2,f_n)}{[W_2(f_1,f_2)]^2}. \tag{3.6}$$

These derivatives would be nil if the generating kernel were of finite rank $L=2$ since the linear dependence of f_1,f_2,f_n then gives a $W_3(f_1,f_2,f_n)=0$. The rhs of (3.5) would then be a constant. If the ratios of Wronskians on the rhs of (3.5) are nearly constant, a reasonable approximant can be defined through

$$(\tilde{A}^{N=2}/f_1)'/(f_2/f_1)'=\text{const}. \tag{3.7}$$

The solution of this differential system yields

$$\tilde{A}^{N=2}(x)=a_1^2f_1(x)+a_2^2f_2(x). \tag{3.8}$$

If the ratios of Wronskians in (3.5) are not constant, a new differentiation and division everywhere gives

$$\frac{[(\tilde{f}/f_1)'/(f_2/f_1)']'}{[(f_3/f_1)'/(f_2/f_1)']'}=\lambda^3+\lambda^4\frac{W_3(f_1,f_2,f_4)}{W_3(f_1,f_2,f_3)}+\cdots+\lambda^n\frac{W_3(f_1,f_2,f_n)}{W_3(f_1,f_2,f_3)}\cdots. \tag{3.9}$$

The derivatives of the ratios of Wronskians on the rhs are

$$\left[\frac{W_3(f_1,f_2,f_n)}{W_3(f_1,f_2,f_3)}\right]'=\frac{W_2(f_1,f_2)\cdot W_4(f_1,f_2,f_3,f_n)}{[W_3(f_1,f_2,f_3)]^2}. \tag{3.10}$$

(see Ref. 6 where the author has given general composition rules for Wronskians of Wronskians). These derivatives would be nil for finite rank kernel $L=3$ and the rhs of (3.9) would then be a constant. Again if these ratios of Wronskians in (3.9) are only approximately constant, an approximant can be defined through

$$\frac{[(\tilde{A}^3/f_1)'/(f_2/f_1)']'}{[(f_3/f_1)'/(f_2/f_1)']'}=\text{const} \tag{3.11}$$

and then

$$\tilde{A}^3=a_1^3f_1(x)+a_2^3f_2(x)+a_3^3f_3(x). \tag{3.12}$$

Ideally this process of alternate differentiation and division can be repeated until one achieves a practical constancy for the ratios of Wronskians of order N, thereby indicating that a linear combination of N Born terms is sufficient. In practice, of course, these tests of constancy are limited by the actual number of Born terms available. If we are dealing with a completely continuous generating operator, it will be ever more closely approximated, and so will its Born terms, by finite rank kernels of ever higher rank. This should manifest itself by an "improved" constancy after each repeated process of alternate differentiation and division. To the same extent an approximate solution will be given by a linear combination of Born terms. We have again taken our cue from completely continuous operators. It should be obvious however that any formal power series, which exhibits a similar tendency towards constant ratios of Wronskians of Born terms, enables us to define heuristically an approximant of the type (3.11).

One may wonder why we have not written the lhs of (3.9) and (3.11) also as ratios of Wronskians which they are. The reason is twofold: First, because it is easiest to show in the form of ratios of derivatives that our test is independent of the parametrization "x" of the Born terms, we can take them as ratios of total derivatives; secondly, if the Born terms have coinciding zeros, as may happen at a threshold, we see that the test is still well defined.

In Sec. 2 we went to some trouble to discredit the use of the coupling constant or expansion parameter λ when it is large, and we have simply used the Born terms to construct approximants which are solutions of differential equations. We will need extra information to fix the constants a_i^N of (3.1). This, as with any differential equation system, may come from boundary conditions or more likely from initial conditions. These initial conditions imply full knowledge of the solution at one point. In theoretical cases this may happen for example in potential theory where some problems are completely solvable say for some energy $k=0$ but not for other values of k.[7] If, for these other values of the energy, perturbatives terms are computed by iteration, our approximant is immediately applicable. In experimental circumstances, determination of the coupling constant itself requires knowledge of the solution at one point, so that this knowledge could be used directly instead to provide initial conditions. With this approach one might even completely bypass the notion of coupling constant and consider it a mere formal device whenever it turns out to be large.

Ideally, if we know the value of the true solution at one point and its derivatives, we can arrange a match between the approximant and the solution at that point. By posing then $A^N(x_0)=f(x_0)$, $A^{N\prime}(x_0)=f'(x_0)$, etc., or equivalently for $\tilde{A}^N(x_0)=\tilde{f}(x_0)$, $\tilde{A}^{N\prime}(x_0)=\tilde{f}'(x_0)$, etc., the coefficients of 3.1 are given as

$$a_i^N=\frac{W_N(f_1,\ldots,f_{i-1},\tilde{f}_i,f_{i+1},\ldots,f_N)}{W_N(f_1,\ldots,f_{i-1},f_i,f_{i+1},\ldots,f_N)}\Bigg|_{x_0}. \tag{3.13}$$

It is well to note at this stage that the proposed approximant is not equivalent to building a kernel of finite rank approximation with

$$k^N(x,x')=\sum_{i=1}^{N} f_i(x)\beta_i(x'), \tag{3.14}$$

of unspecified $\beta_i(x)$. A rapid glance at the "x" dependence of both approximants through the use of N Born terms might give this impression. Indeed the two approximants satisfy the same type of differential equations [e.g., Eq. (3.11)] as we have emphasized at the beginning of this section. The important feature of a Born–Neumann series seemed to us this differential equation rather than the λ dependence of the coefficients $c_i^N(\lambda)$ as explained in Sec. 2[see Eq. (2.16)]. Having made the passage from one approximant to the other via the common differential equation, it might be useful to examine the difference in the λ dependence of their respective coefficients. First, a kernel of finite rank approximation would necessitate ad hoc reconstruction of the $\beta_i(x)$ or $2N$ Born terms to form an equivalent $[N/N]$ Padé, and at any rate the coefficients of the Born terms will be tied to a specific form of λ dependence, i.e., the $c_i^N(\lambda)$ of Eq. (2.13). On the other hand our approximant, with the coefficients defined in Eq. (3.13), uses implicitly the actual form of λ dependence found in the "true" solution at the point x_0. It does not impose a specific λ dependence but seeks rather that of the actual solution, through knowledge of $f(x_0),\ldots,f^{(N-1)}(x_0)$, this λ dependence is in general very different from that of kernel of finite rank approximation.

There is, however, one instance when the two types of approximation coincide, that is, if it happens that the Born series at hand itself stems from an integral equation with a kernel of finite rank N. Then our N approximant and the true solution satisfy exactly the same differential equation as well as having the same initial conditions. Our proposed approximant is then exact for all x and λ. This and other characteristics are studied in detail in the error analysis that follows.

4. ERROR ANALYSIS

With this choice for the coefficients it is interesting to study the structure of the error at points $x\neq x_0$. We define

$$\epsilon^N(x;x_0)\equiv f(x)-A^N(x;x_0)=\tilde{f}(x)-\tilde{A}^N(x;x_0). \tag{4.1}$$

Thus

$$\epsilon^N(x;x_0)= \frac{\tilde{f}(x)W_N(f_1,\ldots,f_N)_{x_0}-f_1(x)W_N(\tilde{f},f_2,\ldots,f_N)_{x_0}\cdots-f_N(x)W_N(f_1,\ldots,f_{N-1},\tilde{f})_{x_0}}{W_N(f_1,\ldots,f_N)_{x_0}} \tag{4.2}$$

and after some manipulations:

$$\epsilon^N(x;x_0)=[W_N(f_1,\ldots,f_N)_{x_0}]^{-1}\cdot\begin{vmatrix} f_1(x_0) & \cdots & f_N(x_0) & \tilde{f}(x_0) \\ f_1'(x_0) & \cdots & f_N'(x_0) & \tilde{f}'(x_0) \\ \cdot & & \cdot & \cdot \\ \cdot & & \cdot & \cdot \\ \cdot & & \cdot & \cdot \\ f_1^{(N-1)}(x_0) & \cdots & f_N^{(N-1)}(x_0) & \tilde{f}^{(N-1)}(x_0) \\ f_1(x) & \cdots & f_N(x) & \tilde{f}(x) \end{vmatrix}. \tag{4.3}$$

This determinental form is convenient to study the error.

(a) If each term of the last row is developed about point x_0 as

$$f_i(x)=f_i(x_0)+f_i'(x_0)(x-x_0)+\cdots+f_i^{(N)}(\xi_i)(x-x_0)^N/N!\,, \tag{4.4}$$

then multiplying each of the first N rows by the relevant coefficients $(x-x_0)^j/j!$ before subtracting from the last row we obtain

$$\epsilon^N(x;x_0)=[W_N(f_1,\ldots,f_N)_{x_0}]^{-1}\begin{vmatrix} f_1(x_0) & \cdots & f_N(x_0) & \tilde{f}(x_0) \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ f_1^{(N-1)}(x_0) & \cdots & f_N^{(N-1)}(x_0) & \tilde{f}^{(N-1)}(x_0) \\ f_1^{(N)}(\xi_1) & \cdots & f_N^{(N)}(\xi_N) & \tilde{f}^{(N)}(\xi) \end{vmatrix}\frac{(x-x_0)^N}{N!}\,, \tag{4.5}$$

where

$$x_0\le\xi_1,\ldots,\xi_N,\xi\le x$$

or equivalently

$$\epsilon^N(x;x_0)=[W_N(f_1,\ldots,f_N)_{x_0}]^{-1}[W_{N+1}(f_1,\ldots,f_N,\tilde{f})_{x_0}(x-x_0)^N/N!+O(x-x_0)^{N+1}]. \tag{4.6}$$

That the error should be proportional to a term $(x-x_0)^N$ comes as no surprise since we chose the coefficients a_i^N using a fit with the solution and its derivatives through $(N-1)$ at x_0.

Further if our ansatz is justified in approximating the solution $\tilde{f}(x)$ by a sum of Born terms, it is useful to write for the solution $f(x)$ an exact relation:

$$\tilde{f}(x)\equiv f(x)-g(x)=\sum_{i=1}^{N}e_i f_i(x)+r(x), \tag{4.7}$$

where $r(x)$ is the x dependence of the solution not reducible to a linear combination of the N Born terms written. Note that we do not know the value of the exact coefficients e_i surely different from the a_i^N. No matter though, since our approximant is self-correcting in this respect. Indeed as the determinant in (4.3) shows, the first N columns remove any linear dependence in the $f_i(x)$ from the last column. Combined with the Taylor expansion as in (4.4) this gives

$$\epsilon^N(x;x_0)=[W_N(f_1,\ldots,f_N)_{x_0}]^{-1}\begin{vmatrix} f_1(x_0) & f_N(x_0) & r(x_0) \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ f_1^{(N-1)}(x_0) & \cdots f_N^{(N-1)}(x_0) & r^{(N-1)}(x_0) \\ f_1^{(N)}(\xi_1) & \cdots f_N^{(N)}(\xi_N) & r^{(N)}(\xi) \end{vmatrix}\frac{(x-x_0)^N}{N!}\,. \tag{4.8}$$

This determinant form for the error shows that the approximant has the advantage of a derivative fitting, as expected, together with the elimination of whatever forms of "x" dependence that happen to be in the N Born terms.

There is yet another advantage, normally one can always write a solution in the form:

$$\tilde{f}(x)=\sum_{i=1}^{N}\lambda^i f_i(x)+\lambda^{N+1}K^{N+1}f(x). \tag{4.9}$$

If this expression is substituted in the last column of (4.3) for the solution and its derivatives either at x_0 or at x, all the terms of the above sum $\sum_{i=1}^{N}$ are eliminated and the error can be written as

$$\epsilon^N(x;x_0)=\lambda^{N+1}[W_N(f_1,\ldots,f_N)_{x_0}]^{-1}\begin{vmatrix} f_1(x_0) & \cdots f_N(x_0) & \mathrm{K}^{N+1}f(x_0) \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ f_1^{(N-1)}(x_0) & \cdots f_N^{(N-1)}(x_0) & (\mathrm{K}^{N+1}f)^{(N-1)}(x_0) \\ f_1(x) & \cdots f_N(x) & \mathrm{K}^{N+1}f(x) \end{vmatrix} . \tag{4.10}$$

In other words for λ small the error is of the $O(\lambda^{N+1})$ just as with a Padé approximant using N Born terms. This result is welcomed, of course, and it does not jeopardize either of the two error reduction features previously described. The Taylor expansion property through row subtraction from the last one still holds as in (4.5). Similarly the column combinations described in (4.8) still reduce the terms of the last column linearly dependent in their x behavior on the Born terms. This is so because the last column, and the N Born terms all result from the application of the same operator K on some function; thus they are linearly dependent in their x behavior to the extent that the kernel of K can be approximated by a kernel of finite rank N.

5. CONCLUSION

The approximant we have proposed is based on considerations of finite rank approximations to completely continuous operators as generators of Born–Neumann series and on the differential equations they satisfy. Retaining these differential equations as the essential features, we find solutions that appear as linear combinations of Born terms. Since the weights of the Born terms tend to become independent of the expansion parameter λ as it gets large for the kernel of finite rank approximations, we have also abolished the hierarchy among Born terms, seeking to determine those weights phenomenologically. New information, often of an empirical nature, will have to be supplied to determine the importance of each "perturbative" term. This may be the main difficulty in practice, and we have discussed in detail coefficients determined by initial conditions.

Among the advantages of the proposed approximant, we note:

(i) It will give the exact solution to a problem of finite rank kernel N with only N Born terms, whereas the Padé approximant requires $2N$.[5]

(ii) Contrary to Padé approximant our proposed approximant will not introduce stray x dependence in the denominators and will thus preserve all the analytical properties in the variables covered by "x".

(iii) The linear combination of Born terms preserves such an important property as crossing symmetry if one is thinking of high energy applications. Its drawbacks are

(i) Required knowledge of the solution and $(N-1)$ derivatives at one point.

(ii) Unitarity fails in general and has to be examined specifically.

(iii) It cannot obviously predict poles as function of λ, since λ has been excluded from our considerations. It could, however, "witness" a pole, through the coefficients determined at x_0, if the position of this pole in λ is completely independent of the value of x. This is indeed the situation for Fredholm resolvents whose pole positions in λ are independent of x. Our approximant is thus better suited for approximating functions free of poles arising through a variable coupling strength. The proposed approximant has the characteristics of a best fit to the true solution in the following sense. One maintains simultaneously three features since the error is

(a) like that of a Taylor fit at the point x_0, that is, proportional to $(x-x_0)^N$, useful for x close to x_0,

(b) like that of a Padé approximant constructed out of N Born terms, that is, of order $O(\lambda^{N+1})$, useful strictly speaking for small λ only,

(c) proportional to a determinant which tends to vanish in so far as the "x-shapes" present in the N Born terms are sufficient to describe by linear combination the x dependence of the solution.

Finally we recall that the weights of the Born terms, i.e., the coefficients a_i^N, are obtained independently of the choice of parametrization for the x variables since the ratios of Wronskians involved are made up of successive ratios of derivatives.

[1]P.R. Graves-Morris, *Padé Approximants* (The Institute of Physics, London and Bristol, 1973).
[2]P.R. Graves-Morris, *Padé Approxmants and Their Applications* (Academic, New York, 1973).
[3]G.A. Baker and J.L. Gammel, *Padé Approximants in Theoretical Physics* (Academic, New York, 1971).
[4]F.W. Byron and R.W. Fuller, *Mathematics of Classic and Quantum Physics* (Addison-Wesley, Reading, Mass., 1970), Vol. II, p. 575.
[5]R. Chisolm, J. Math. Phys. 4, 1506 (1963).
[6]R.C. Brunet, J. Math. Phys. 16, 1112 (1975).
[7]W.M. Frank, D.J. Land, and R.M. Spector, Rev. Mod. Phys. 43, 36 (1971).

A multipoint interpolation method based on variational principles for functionals of the solution to linear equations

Edward T. Cheng and Robert W. Conn

Nuclear Engineering Department, The University of Wisconsin, Madison, Wisconsin 53706
(Received 28 May 1975)

A method is derived for using variational expressions to interpolate among known values of a functional of the solution to linear equations. For linear functionals of the solution to an inhomogeneous equation, the interpolation expression is exact at N distinct points when N distinct functions are used, each of which is the solution of the underlying Euler equation. Two point variational interpolation is derived to interpolate on the value of an eigenvalue using the Rayleigh quotient. Illustrative examples are given based on neutron transport studies of fusion reactor blanket systems and applications to sensitivity and optimization studies in reactor theory are discussed.

I. INTRODUCTION

Variational theory has been widely used in mathematical physics to evaluate functionals or to derive approximate theories. In the former application, the motivation for using variational techniques is the fact that errors are second order with respect to inaccuracies in trial functions. An additional motivation is that one is often actually interested in a functional of the solution to an equation describing a physical system, rather than in the solution itself. Examples are the evaluation of transport coefficients for gases and plasmas and the evaluation of various scalar products of the neutron or gamma flux in fission and fusion reactor neutronics studies. In general, it is of interest to estimate inner products of the form

$$(S^\dagger, \phi), \tag{1}$$

where ϕ satisfies a linear inhomogeneous equation

$$L\phi = S. \tag{2}$$

The adjoint equation is

$$L^\dagger \phi^\dagger = S^\dagger. \tag{3}$$

In neutron transport theory, L is the Boltzmann transport operator,[1] and S is a source.

Two widely used variational principles to estimate linear functionals of the solution to an inhomogeneous equation are the Schwinger[2] and Roussopoulos[3] variational principles. Both these principles have been generalized by Pomraning[4] to provide an estimate of an arbitrary functional, rather than just a linear one. Several recent papers have also treated the problem of estimating changes in a functional of interest[5,6] using variational forms accurate to second order in the change.[7]

In this paper, a method for using variational expressions to interpolate among known values of a given function is derived. The linear operator, $L(\alpha)$, is assumed to depend in some known way on a set of parameters, α. To estimate the effect of changes in α on the functional of interest, e.g., $(S^\dagger, \phi)$, the standard procedure has been to let $\alpha = \alpha_1$ be defined as a reference system and to use $L(\alpha_1)$ and $L^\dagger(\alpha_1)$ to determine solutions ϕ_1 and $\phi_1^\dagger$. Then ϕ_1 and $\phi_1^\dagger$ are used as trial functions in either the Schwinger or Roussopoulos functionals to assess the effect of changing α on the response functional of interest. Often, the perturbation introduced is large and/or more than one reference system is appropriate. In such cases, the method of variational interpolation to be described here can be used to interpolate among several reference values. For linear functionals, an expression is derived which is exact at an arbitrary number of reference points and which can be used to interpolate among them. Two point variational interpolation is derived to interpolate on the value of an eigenvalue using the Rayleigh quotient. Some illustrative numerical examples are given based on neutron transport studies of fusion reactor blanket systems which have recently become of greatly increased interest.[8]

II. THEORY OF VARIATIONAL INTERPOLATION

A. Linear functionals and inhomogeneous equations

The simpliest illustration of the basic idea is to consider two point variational interpolation for linear functionals of the solution to a linear inhomogeneous equation. Letting $(S^\dagger, \phi)$ be the linear functional of interest, the Roussopoulos functional,

$$I_R[\phi^\dagger, \phi; \alpha] = (S^\dagger(\alpha), \phi) + (\phi^\dagger, S(\alpha) - L(\alpha)\phi) \tag{4}$$

is stationary about the exact value of $(S^\dagger(\alpha), \phi)$ with Euler equations

$$L(\alpha)\phi = S(\alpha), \tag{5}$$

$$L^\dagger(\alpha)\phi^\dagger = S^\dagger(\alpha). \tag{6}$$

As noted, α represents parameters in the operator L (for example, cross sections or densities when L is the Boltzmann transport operator) and S and $S^\dagger$ may depend on α. Let us now characterize two reference systems by the parameters α_1 and α_2. To estimate $(S^\dagger(\alpha), \phi)$ at a point α not the same as α_1 or α_2, we chose trial functions ϕ_1 and $\phi_2^\dagger$ which satisfy, respectively,

$$L(\alpha_1)\phi_1 = S(\alpha_1), \tag{7}$$

$$L^\dagger(\alpha_2)\phi_2^\dagger = S^\dagger(\alpha_2). \tag{8}$$

The simpliest form of the method of variational interpolation follows from noting that the functional

$$I_R[\phi_2^\dagger, \phi_1; \alpha] = (S^\dagger(\alpha), \phi_1) + (\phi_2^\dagger, S(\alpha) - L(\alpha)\phi_1) \tag{9}$$

is exact at both reference points. Clearly, for $\alpha = \alpha_1$, the operator L is $L(\alpha_1)$ and the source is $S(\alpha_1)$.

Thus, the second term in Eq. (9) is zero and $I_R[\phi_2^\dagger, \phi_1; \alpha_1]$ is exact. For a system with $\alpha = \alpha_2$, where $L = L(\alpha_2)$, $S = S(\alpha_2)$, and $S^\dagger = S^\dagger(\alpha_2)$, we use

$$(\phi_2^\dagger, L(\alpha_2)\phi_1) = (S^\dagger(\alpha_2), \phi_1), \tag{10}$$

from which it follows that $I_R[\phi_2^\dagger, \phi_1; \alpha_2]$ is also exact. Thus, the functional $I_R[\phi_2^\dagger, \phi_1; \alpha]$ can be used to interpolate in α and thus estimate other values of the basic functional.

The Schwinger functional,

$$I_s[\phi^\dagger, \phi, \alpha] = (S^\dagger(\alpha), \phi)(\phi^\dagger, S(\alpha))/(\phi^\dagger, L(\alpha)\phi) \tag{11}$$

is also exact at $\alpha = \alpha_1$ and $\alpha = \alpha_2$ when $\phi = \phi_1$ and $\phi^\dagger = \phi_2^\dagger$ are used as input functions. The proof is equally straightforward. At $\alpha = \alpha_1$, use Eq. (7) to show that $I_s[\phi_2^\dagger, \phi_1; \alpha_1] = (S^\dagger(\alpha_1), \phi_1)$. At $\alpha = \alpha_2$, again use Eq. (10) to find $I_s[\phi_2^\dagger, \phi_1; \alpha_2] = (\phi_2^\dagger, S(\alpha_2))$, which, of course, is equal to $(S^\dagger(\alpha_2), \phi_2)$. Thus, the Schwinger functional can equally well be employed to interpolate in α and it can have advantages over the Roussopoulos functional, as has been discussed recently.[9] We will expand on this shortly.

A three point interpolation formula can readily be derived, and it suggests the procedure to follow in constructing a general proof. Consider three reference systems, α_1, α_2, and α_3, and the trial functions

$$\phi_T(\alpha) = \phi_1 + a(\alpha)(\phi_2 - \phi_1), \tag{12}$$

$$\phi_T^\dagger(\alpha) = b(\alpha)\phi_3^\dagger, \tag{13}$$

where ϕ_1, ϕ_2, and $\phi_3^\dagger$ are solutions to the appropriate equations for the subscripted reference points. Insert $\phi_T(\alpha)$ and $\phi_T^\dagger(\alpha)$ into $I_R(\phi_T^\dagger, \phi_T; \alpha)$ and solve the equations $\partial I_R/\partial a = 0$, $\partial I_R/\partial b = 0$. This yields

$$a(\alpha) = (\phi_3^\dagger, S(\alpha) - L(\alpha)\phi_1)/(\phi_3^\dagger, L(\alpha)(\phi_2 - \phi_1)), \tag{14}$$

$$b(\alpha) = (S^\dagger(\alpha), \phi_2 - \phi_1)/(\phi_3^\dagger, L(\alpha)(\phi_2 - \phi_1)). \tag{15}$$

Using these expressions in ϕ_T and $\phi_T^\dagger$ gives the functional

$$I_3[\phi_T^\dagger, \phi_T; \alpha] = (S^\dagger(\alpha), \phi_1) + (\phi_3^\dagger, S(\alpha) - L(\alpha)\phi_1)(S^\dagger(\alpha), \phi_2 - \phi_1)/(\phi_3^\dagger, L(\alpha)(\phi_2 - \phi_1)), \tag{16}$$

which is exact when α equals α_1, α_2, or α_3.

A general proof can be constructed for N forward trial functions and $N-1$, N, or $N+1$ distinct adjoint trial functions (or N adjoint functions and $N-1$, N, or $N+1$ distinct forward functions.) The proof for $2N$ distinct reference systems proceeds as follows. Let

$$\phi_T(\alpha) = \phi_1 + \sum_{i=2}^{N} a_i(\alpha)(\phi_i - \phi_1) \tag{17}$$

and

$$\phi_T^\dagger(\alpha) = \phi_{N+1}^\dagger + \sum_{i=N+2}^{2N} b_i(\alpha)(\phi_i^\dagger - \phi_{N+1}^\dagger), \tag{18}$$

where the ϕ_i satisfy

$$L(\alpha_i)\phi_i = S(\alpha_i) \tag{19}$$

and the $\phi_i^\dagger$ satisfy

$$L^\dagger(\alpha_i)\phi_i^\dagger = S^\dagger(\alpha_i). \tag{20}$$

All α_i are distinct and the indices can clearly be arbitrarily assigned. Inserting Eqs. (17) and (18) into $I_R[\phi_T^\dagger, \phi_T; \alpha]$ and carrying out a Rayleigh–Ritz procedure, $(\partial I_R/\partial a_i = 0, \partial I_R/\partial b_i = 0)$, the following set of coupled algebraic equations are obtained:

for the coefficients $a_i(\alpha)$,

$$\begin{aligned} &a_2(\alpha)((\phi_i^\dagger - \phi_{N+1}^\dagger), L(\alpha)(\phi_2 - \phi_1)) \\ &\quad + a_3(\alpha)((\phi_i^\dagger - \phi_{N+1}^\dagger), L(\alpha)(\phi_3 - \phi_1)) + \cdots \\ &\quad + a_N(\alpha)((\phi_i^\dagger - \phi_{N+1}^\dagger), L(\alpha)(\phi_N - \phi_1)) \\ &= ((\phi_i^\dagger - \phi_{N+1}^\dagger), S(\alpha) - L(\alpha)\phi_1), \\ &\qquad i = N+2, N+3, \ldots, 2N; \end{aligned} \tag{21}$$

for the coefficients $b_i(\alpha)$,

$$\begin{aligned} &b_{N+2}(L^\dagger(\alpha)(\phi_{N+2}^\dagger - \phi_{N+1}^\dagger), (\phi_i - \phi_1)) \\ &\quad + b_{N+3}(L^\dagger(\alpha)(\phi_{N+3}^\dagger - \phi_{N+1}^\dagger), (\phi_i - \phi_1)) + \cdots \\ &\quad + b_{2N}(L^\dagger(\alpha)(\phi_{2N}^\dagger - \phi_{N+1}^\dagger), (\phi_i - \phi_1)) \\ &= (S^\dagger(\alpha) - L^\dagger(\alpha)\phi_{N+1}^\dagger, (\phi_i - \phi_1)), \\ &\qquad i = 2, 3, \ldots, N. \end{aligned} \tag{22}$$

The functional $I_{2N}[\phi_T^\dagger, \phi_T; \alpha]$ is formed by using Eqs. (17) and (18) as trial functions with coefficients $\{a_i(\alpha)\}$ and $\{b_i(\alpha)\}$ determined by solving Eqs. (21) and (22). Proving that $I_{2N}[\phi_T^\dagger, \phi_T; \alpha]$ is exact whenever $\alpha = \alpha_i$, $i = 1, 2, \ldots, 2N$, is equivalent to proving that, at $\alpha = \alpha_p$, one of the trial functions equals the exact solution, i.e., either $\phi_T = \phi_p$ or $\phi_T^\dagger = \phi_p^\dagger$.

For $\alpha = \alpha_1$ the right-hand side of Eq. (21) vanishes. Since the coefficients are linearly independent (the α_i are distinct), it follows that $a_i(\alpha_1) = 0$ for $i = 2, 3, \ldots, N$. Thus, $\phi_T(\alpha_1) = \phi_1$ and $I_{2N}[\phi_T^\dagger, \phi_T; \alpha_1]$ is exact. For $\alpha = \alpha_k \neq \alpha_1$, rewrite the right-hand side of Eq. (21) as

$$((\phi_i^\dagger - \phi_{N+1}^\dagger), S(\alpha_k) - L(\alpha_k)\phi_1) = ((\phi_i^\dagger - \phi_{N+1}^\dagger, L(\alpha_k)(\phi_k - \phi_1)). \tag{23}$$

Then the coupled algebraic equations become

$$\begin{aligned} &a_2((\phi_i^\dagger - \phi_{N+1}^\dagger), L(\alpha_k)(\phi_2 - \phi_1)) + \cdots \\ &\quad + (a_k - 1)((\phi_i^\dagger - \phi_{N+1}^\dagger), L(\alpha_k)(\phi_k - \phi_1)) + \cdots \\ &\quad + a_N((\phi_i^\dagger - \phi_{N+1}^\dagger), L(\alpha_k)(\phi_N - \phi_1)) = 0. \end{aligned} \tag{24}$$

In this form, we see that the linear independence of the inner product coefficients again implies

$$a_2 = a_3 = \cdots = a_k - 1 = \cdots = a_N = 0. \tag{25}$$

Thus, $a_i = \delta_{ik}$, $\phi_T(\alpha_k) = \phi_k$, and this guarantees that $I_{2N}[\phi_T^\dagger, \phi_T; \alpha_k]$ is exact for any $k = 2, 3, \ldots, N$. Therefore, we have proven that $I_{2N}[\phi_T^\dagger, \phi_T; \alpha]$ is exact whenever $\alpha = \alpha_1, \alpha_2, \ldots, \alpha_N$.

Similar arguments applied to Eq. (22) prove that for $\alpha = \alpha_k$, $k = N+1, N+2, \ldots, 2N$, $I_{2N}[\phi_T^\dagger, \phi_T; \alpha_k]$ is also exact. Thus, $I_{2N}[\phi_T^\dagger, \phi_T; \alpha]$ involves $2N$ distinct trial functions and takes on the exact value for the corresponding $2N$ distinct reference systems. This functional can therefore be used to interpolate in α among these exact values and constitutes a multipoint variational interpolation. A proof for N reference ϕ functions and $N-1$ or $N+1$ distinct reference adjoint functions can be carried through following the same procedure.

The form of the error term in variational interpolation can be illustrated by examining the two point formula. Again consider the linear functional, $R(\alpha)=(S^\dagger(\alpha),\phi)$, of the solution of a linear inhomogeneous equation. When an altered $S^\dagger=S_T^\dagger$ and an altered $\phi=\phi_T$ are used, the first order change relative to the reference point in R is given by

$$\delta R=(\delta S^\dagger,\phi)+(S^\dagger,\delta\phi), \tag{26}$$

where $\delta S^\dagger=S_T^\dagger-S^\dagger$ and $\delta\phi=\phi_T-\phi$. This expression is the sum of an error term due to the perturbation itself, which changes $S^\dagger$, and the error induced because the perturbation in turn effects the solution. The change, δR, can be rewritten using $L\delta\phi=-\delta L\phi+\delta S$, where the operator L_T has been written as $L+\delta L$ and S_T as $S+\delta S$. δR becomes

$$\delta R=(\delta S^\dagger,\phi)+(\phi^\dagger,\delta S)-(\phi^\dagger,\delta L\phi). \tag{27}$$

Assume δS, $\delta S^\dagger$, and δL depend linearly on the change in a parameter α so that

$$\delta S=s\delta\alpha, \tag{28a}$$

$$\delta S^\dagger=s^\dagger\delta\alpha, \tag{28b}$$

and

$$\delta L=H\delta\alpha. \tag{29}$$

Then the derivative of R with respect to α at the reference value α_1 is

$$\left.\frac{\delta R}{\delta\alpha}\right|_{\alpha_1}=(s^\dagger,\phi_1)+(\phi_1^\dagger,s)-(\phi_1^\dagger,H\phi_1). \tag{30}$$

It is easily shown that both the Roussopoulos functional, Eq. (4), and the Schwinger functional, Eq. (11), preserve the exact slope at $\alpha=\alpha_1$ if trial functions ϕ_1 and $\phi_1^\dagger$ are used.

In the method of variational interpolation, the trial functions are taken at distinct reference points, for example, ϕ_1 at $\alpha=\alpha_1$ and $\phi_2^\dagger$ at $\alpha=\alpha_2$. The slope of the functional does not, however, preserve the exact slope at either α_1 or α_2. Indeed, for two point interpolation where the changes depend linearly on α, the Roussopoulos form, Eq. (9) yields a straight line interpolation between $(S^\dagger(\alpha_1),\phi_1)$ and $(S^\dagger(\alpha_2),\phi_2)$. The difference between the slope using Eq. (9) and the exact slope is $(\delta\phi_{21}^\dagger,\ H\phi_1-s)$, where $\delta\phi_{21}^\dagger=\phi_2^\dagger-\phi_1^\dagger$.

Variational interpolation using the Schwinger functional comes closer to preserving the slope. By using Eq. (11) with ϕ_1 and $\phi_2^\dagger$ as input functions, the approximate slope is

$$\left.\frac{\partial I_s}{\partial\alpha}\right|_{\alpha_1}=(s^\dagger,\phi_1)-\frac{(S^\dagger(\alpha_1),\phi_1)}{(\phi_2^\dagger,S)}\left[(\phi_2^\dagger,H\phi_1)-(\phi_2^\dagger,s)\right]. \tag{31}$$

Let $\Delta(\partial R/\partial\alpha)$ be the difference in slope from the exact value. One then finds that

$$\Delta\left(\frac{\partial R}{\partial\alpha}\right)=\left(\left(\delta\phi_{21}^\dagger-\frac{(\delta\phi_{21}^\dagger,S)}{(\phi_1^\dagger,S)}\,\phi_1^\dagger\right),(H\phi_1-s)\right), \tag{32}$$

neglecting second order terms. Compared with $(\delta\phi_{21}^\dagger,\ H\phi_1-s)$, we see now the additional term $[(\delta\phi_{21}^\dagger,s)/(\phi_1^\dagger,S)]\phi_1^\dagger$ in the inner product on the right-hand side of Eq. (32). This term is independent of the amplitude of $\phi_1^\dagger$ but does depend on the difference $\delta\phi_{21}^\dagger$. This added term attempts to correct for first order differences in the adjoint functions. This, if the shape of the adjoint function tends to be preserved, the slope will tend to be preserved to second order. Further, interpolation between α_1 and α_2 based on Eq. (11) will not be linear in α. This is an important distinction between the Roussopoulos and Schwinger functionals which will be clearly illustrated in the numerical examples. For higher order interpolation, the method used to derive the combining coefficients, $\{a_i\}$ and $\{b_i\}$, is the same as that applied to derive the Schwinger principle from the Roussopoulos functional.[10] The interpolation will therefore be nonlinear and should have the same renormalized character as the Schwinger principle.[9] Moreover, if one chooses ϕ_i and $\phi_i^\dagger$ at the same reference point in Eqs. (17) and (18), i.e.,

$$\phi_T(\alpha)=\phi_1+\sum_{i=2}^{N}a_i(\alpha)(\phi_i-\phi_1), \tag{17}$$

$$\phi_T^\dagger(\alpha)=\phi_1^\dagger+\sum_{i=2}^{N}b_i(\alpha)(\phi_i^\dagger-\phi_1^\dagger), \tag{18'}$$

the general procedure is equivalent to the variational synthesis method discussed by Kaplan.[11]

Now consider the exact form of the error term in two point variational interpolation when ϕ_1 and $\phi_1^\dagger$, evaluated at $\alpha=\alpha_1$, are used as trial functions in the Roussopoulos principle, Eq. (4), to estimate $(S^\dagger(\alpha),\phi)$, $\alpha\neq\alpha_1$. The error is

$$\epsilon_R=-(\delta\phi_1^\dagger,L(\alpha)\delta\phi_1)+(\delta\phi_1^\dagger,\delta S). \tag{33}$$

Here, $\delta S=S-S_1$, $\delta\phi_1=\phi-\phi_1$, $\delta\phi_1^\dagger=\phi^\dagger-\phi_1^\dagger$, and ϕ and $\phi^\dagger$ are the exact solutions in system α. When ϕ_1 and $\phi_2^\dagger$ are used in the variational interpolation method, the error is

$$\epsilon_{VI}=-(\delta\phi_2^\dagger,L(\alpha)\delta\phi_1)+(\delta\phi_2^\dagger,\delta S). \tag{34}$$

[This is another way of proving Eq. (9) is exact at $\alpha=\alpha_1$ and $\alpha=\alpha_2$.] By comparing Eqs. (33) and (34), it becomes clear that variational interpolation relies on cancellation of error for α between α_1 and α_2. That is, e.g., as α approaches α_1 from α_2, $\delta\phi_2^\dagger$ is increasing while $\delta\phi_1$ and δS are tending to zero. Thus, one should not expect great accuracy if α does not lie between the two reference parameters. Similar error terms can be derived for higher order interpolation formulas, and they all show the same basic characteristic of cancellation of error.

B. Two point variational interpolation and homogeneous equations

The Rayleigh quotient[4] is a homogeneous functional which is widely used to estimate eigenvalues. Because an eigenvalue equation is homogeneous, only homogeneous functional can be of interest. As such, these functionals are nonlinear and we have not succeeded in constructing N point variational interpolation procedures in this case. Indeed, because the functionals of interest here are nonlinear, it may not be possible to do so. It is possible, however, to construct the simpliest case, two point variational interpolation, and this can sometimes be useful. For example, it is often of

interest to determine the sensitivity of an energy level to the interaction potential in quantum mechanics. If the potential is characterized by one or more parameters which can vary over a specified range, interpolation can be used to determine the change in the eigenvalue as the parameters change.

Consider, therefore, the general eigenvalue equation

$$L(\alpha)\phi_i = \lambda_i F(\alpha)\phi_i \tag{35}$$

and the adjoint equation

$$L^\dagger(\alpha)\phi_i^\dagger = \lambda_i F^\dagger(\alpha)\phi_i^\dagger. \tag{36}$$

It is assumed that $L(\alpha)$ and $F(\alpha)$ are real, though not necessarily self-adjoint, and that the eigenvalues λ_i are discrete and nondegenerate. ϕ_i and $\phi_i^\dagger$ are biorthogonal with respect to $F(\alpha)$, i.e.,

$$(\phi_j^\dagger, F\phi_i) = (F^\dagger\phi_j^\dagger, \phi_i) = \delta_{ij}. \tag{37}$$

A variational expression for the kth eigenvalue is the Rayleigh quotient

$$E_\lambda[\phi_K^\dagger, \phi_K; \alpha] = (\phi_K^\dagger, L(\alpha)\phi_K)/(\phi_K^\dagger, F(\alpha)\phi_K). \tag{38}$$

Consider two systems characterized by α_1 and α_2. Ordinarily, one evaluates the effect of changes in α from, e.g., α_1 by using ϕ_{1K} and $\phi_{1K}^\dagger$ as trial functions. These functions are solutions of

$$L(\alpha_1)\phi_{1K} = \lambda_K(\alpha_1)\phi F(\alpha_1)\phi_{1K} \tag{39}$$

and

$$L^\dagger(\alpha_1)\phi_{1K}^\dagger = \lambda_K(\alpha_1)F^\dagger(\alpha_1)\phi_{1K}^\dagger, \tag{40}$$

respectively. With these trial functions, one proceeds to use $L(\alpha)$ and $F(\alpha)$ to evaluate the Rayleigh quotient.

The expression for $E_\lambda[\phi_{1K}^\dagger, \phi_{1K}; \alpha]$ can be written as

$$E_\lambda[\phi_{1K}^\dagger, \phi_{1K}; \alpha] = \lambda_K(\alpha)\left(1 + \frac{(\delta\phi_{1K}^\dagger, L(\alpha)\delta\phi_{1K}) - (\delta\phi_{1K}^\dagger, F(\alpha)\delta\phi_{1K})}{(\phi_{1K}^\dagger, F(\alpha)\phi_{1K})}\right), \tag{41}$$

where $\delta\phi_{1K} = \phi_K - \phi_{1K}$, $\delta\phi_K^\dagger = \phi_K^\dagger - \phi_{1K}^\dagger$, and higher order terms have been neglected.

Now consider choosing ϕ_{1K} and $\phi_{2K}^\dagger$ as trial functions, where ϕ_{1K} satisfies Eq. (39) and $\phi_{2K}^\dagger$ satisfies

$$L^\dagger(\alpha_2)\phi_{2K}^\dagger = \lambda_K(\alpha_2)F^\dagger(\alpha_2)\phi_{2K}^\dagger. \tag{42}$$

The functional

$$E_\lambda[\phi_{2K}^\dagger, \phi_{1K}; \alpha] = (\phi_{2K}^\dagger, L(\alpha)\phi_{1K})/(\phi_{2K}^\dagger, F(\alpha)\phi_{1K}) \tag{43}$$

is exact when α equals either α_1 or α_2 and can therefore be used to interpolate for $\lambda_K(\alpha)$ when α differs from from α_1 or α_2. Further, this functional can be expressed, using $\delta\phi_{2K}^\dagger = \phi_K^\dagger - \phi_{2K}^\dagger$, as

$$E_\lambda[\phi_{2K}^\dagger, \phi_{1K}; \alpha] = \lambda_K(\alpha)\left(1 + \frac{(\delta\phi_{2K}^\dagger, L(\alpha)\delta\phi_{1K}) - (\delta\phi_{2K}^\dagger, F(\alpha)\delta\phi_{1K})}{(\phi_{2K}^\dagger, F(\alpha)\phi_{1K})}\right), \tag{44}$$

neglecting higher order terms. The cancellation of error characteristic is again clear on examination of the interpolation formula. As α approaches α_1, $\delta\phi_{2K}^\dagger$ remains finite as $\delta\phi_{1K}$ tends to zero while the reverse is true as α tends to α_2. This is analogous to the result found previously for linear functionals of the solution to an inhomogeneous equation.

III. ILLUSTRATIVE NUMERICAL APPLICATION

To illustrate the application of variational interpolation, we have examined a relevant problem of current interest in the neutron transport analysis of conceptual fusion reactor blanket systems. A quantity of primary interest for a reactor based on fusions of deuterium and tritium is the tritium breeding ratio, i.e., the number of tritons produced in the blanket per triton consumed. The sample blanket is shown in Fig. 1. Tritium is produced by neutron reactions in lithium, particularly the $^6\mathrm{Li}(n,\alpha)t$ and $^7\mathrm{Li}(n,n'\alpha)t$ reactions. We study here the breeding ratio from reactions in ^{6}Li, labeled T_6, from ^{7}Li, labeled T_7, and the total breeding ratio, labeled BR. T_6, T_7, and BR are defined as

$$T_6 = (\Sigma_6(n,\alpha), \phi),$$

$$T_7 = (\Sigma_7(n,n'\alpha), \phi),$$

and

$$\mathrm{BR} = [(\Sigma_6(n,\alpha) + \Sigma_7(n,n'\alpha)), \phi].$$

ϕ is normalized to one incident 14.1 MeV neutron, and $\Sigma_6(n,\alpha)$ and $\Sigma_7(n,n'\alpha)$ are the macroscopic, energy and space dependent cross sections for the two pertinent nuclear reactions. This, for T_6, $S^\dagger = \Sigma_6(n,\alpha)$ while for BR, $S^\dagger = \Sigma_6(n,\alpha) + \Sigma_7(n,n'\alpha)$.

The numerical evaluation of the inner products required for variational interpolation were carried out using the program SWANLAKE,[12] developed to apply conventional variational procedures. The computational method to solve the neutron transport equation in multigroup form and the nuclear data employed are the same as described previously.[8]

The illustrative examples are based on asking the question, "How does the breeding ratio change as a function of the percentage of structural material in the tritium breeding zones?" [Zones (2) and (4).]

We have chosen 5% structure as reference system α_1 and 25% structure as reference system α_2. Two point variational interpolation is used in the analysis. ϕ_1 and $\phi_1^\dagger$ have also been used as trial functions in Eqs. (4) and (11) to provide a comparison with the more conventional application of variational techniques. The parameters

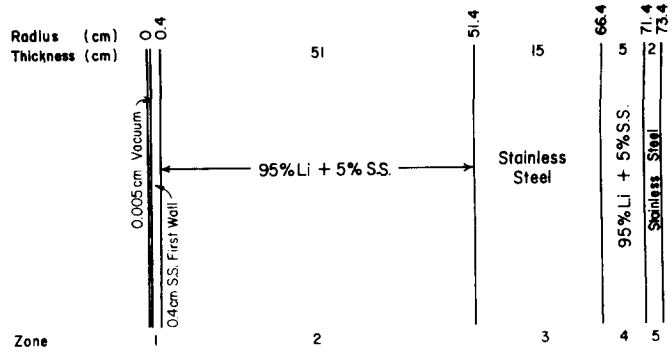

FIG. 1. Schematic of a fusion reactor blanket.

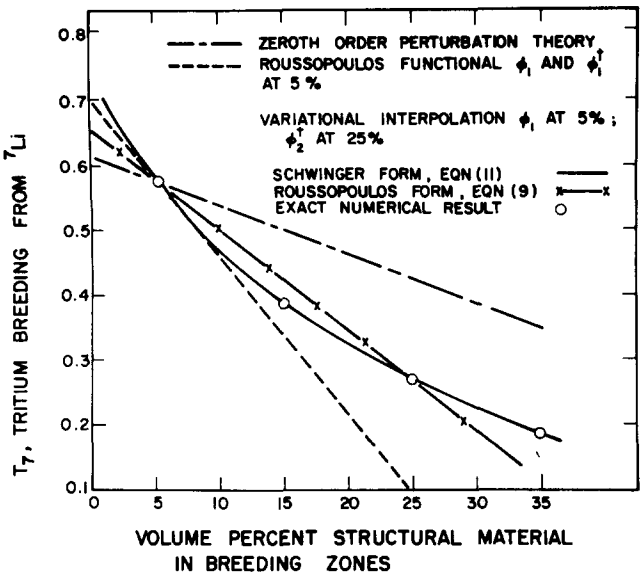

FIG. 2. Variation of T_7 with the amount of structure in the breeding zones (zones 2 and 4 of Fig. 1.)

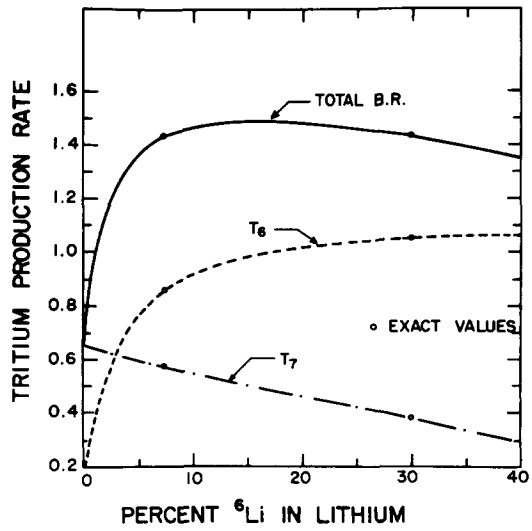

FIG. 3. The results of two point variational interpolation using Eq. (10) with ϕ_1 evaluated at 7.42% ^{6}Li and ϕ_2 evaluated at 30% ^{6}Li.

α used in the theory are the appropriate atomic densities of the materials in zones (2) and (4).

Figure 2 shows T_7 as a function of the percent structure based on several calculational procedures. The open circles are taken as exact from direct numerical calculation. Zeroth order perturbation theory is simply the evaluation of (Σ_7, ϕ_1), where Σ_7 changes as the percentage of lithium changes in the breeding zones. The value of T_7 is preserved at the reference point but not the slope. Two point variational interpolation based on the Roussopoulos functional also does not preserve slope but gives correct values at the two reference points. The Roussopoulos functional using ϕ_1 and $\phi_1^\dagger$ as the trial functions preserves both the value of T_7 and the slope at reference point α_1 but is quite inaccurate at $\alpha = \alpha_2$. Finally, two point variational interpolation based on the Schwinger functional, Eq. (11), is exact when α equals α_1 or α_2 and yields a nonlinear interpolation that is quite close to the exact values for α between α_1 and α_2.

As a second example, the change in BR, T_6, and T_7 as a function of the fraction of ^{6}Li making up the lithium in zones (2) and (4) has been evaluated. (Natural lithium is 7.42% ^{6}Li and 92.58% ^{7}Li.) Two point interpolation has been used with α_1 at 7.42% ^{6}Li and α_2 at 30% ^{6}Li. The results are given in Fig. 3 with the open circles indicating exact values. Again, formula (11) yields a nonlinear interpolation between the reference values.

[1] G.I. Bell. S. Glasstone, *Nuclear Reactor Theory* (Van Nostrand Reinhold, New York, 1970).
[2] R.G. Newton, *Scattering Theory of Particles of Waves* (McGraw-Hill, New York, 1966), p. 319.
[3] P. Roussopoulos, C.R. Acad. Sci. Paris **236**, 1858 (1953).
[4] G.C. Pomraning, J. Soc. Ind. Appl. Math. **13**, 511 (1965).
[5] L.N. Usachev, J. Nucl. Energy, Pts. A/B, **18**, 571 (1964).
[6] A. Gandini, J. Nucl. Energy, Pts. A/B, **21**, 755 (1967).
[7] W.M. Stacey, Jr., J. Math. Phys. **13**, 1119 (1972).
[8] M.A. Abdou and R.W. Conn, Nucl. Sci. Eng. **55**, 256 (1974), and references cited in this paper.
[9] R.W. Conn, Nucl. Sci. Eng. **55**, 468 (1974).
[10] D.S. Selengut, Hanford Laboratories Report, HW-59126 (1959).
[11] S. Kaplan, "Synthesis Methods in Reactor Analysis," in *Advances in Nuclear Science and Technology*, edited by P. Greebler and E. Henly (Academic, New York, 1966), Vol. 3.
[12] D.E. Bartine, F.R. Mynatt, and E.M. Oblow, "SWANLAKE, A Computer Code Utilizing ANISN Radiation Transport Calculations for Cross Section Sensitivity Analysis," Oak Ridge National Lab. Report, ORNL-TM-3809 (May 1973).

Scattering of scalar waves from a Schwarzschild black hole

Norma G. Sanchez*

Department of Theoretical Astrophysics, Paris Observatory, 92190 Meudon, France
(Received 27 May 1975)

The scattering of scalar waves from a Schwarzschild black hole is investigated for wavelengths much less than the gravitational radius (r_s). Explicit expressions for scattering parameters are obtained for two cases: high angular momenta and low angular momenta. In the first case we obtain the phase shifts and absorption coefficient with the JWKB method. The elastic differential cross section and the total absorption cross section are also calculated. For low angular momenta we present a method based in the DWBA (distorted wave Born approximation). With this method, the phase shifts and the absorption coefficients are obtained.

I. INTRODUCTION

The scattering and absorption of scalar waves by a Schwarzschild field is investigated here. This subject has been previously considered by several authors, [1,2] but exact expressions for the phase shifts and for the cross sections has not been found.

In this scattering problem, the choice of the boundary conditions needs special attention. Every solution of the radial wave equation remains bounded on the Schwarzschild radius r_s, and consequently every solution is "physically acceptable."[2] This property, due to the presence of the singular attractive term propertional to $-(r-r_s)^{-2}$ in the effective potential, is intimately connected with the wave capture by the black hole.

The physical solution of the wave equation must be selected in order to have purely ingoing waves on the horizon $r=r_s$.[1]

In this paper we study the scattering problem for waves of short wavelength. That is, wavelength much less than the Schwarzschild radius r_s.

We find approximate expressions for the phase shifts, the absorption coefficient, the elastic and the capture cross sections for two cases:

(a) High angular momenta ($l \gg kr_s$) and low angular momenta ($l \ll kr_s$). In the first case, we find the phase shifts by the JWKB approximation in a partial wave analysis. They are expressed in an expansion in powers of (r_s/b), where the impact parameter $b=(l+1/2)/k$ is large, in this case $b \gg r_s$.

With these phase shifts and by means of the eikonal formalism, an expression for the differential cross section, valid for small angles, is obtained. For $b \sim r_s$, the absorption coefficient for the lth partial wave is calculated. With this partial wave absorption coefficient, the total capture cross section is obtained in Sec. III.

In all of our results one recovers the geometrical optics expressions in the limit $kr_s \to \infty$.

For the case $l \ll kr_s$, we exhibit a method to calculate scattering parameters, based on a modification of the DWBA (Distorted Wave Born Approximation). In this approximation we find the phase shifts and the partial wave absorption coefficients. Here, the choice of the boundary conditions is discussed in Sec. IV.

By using the Schwarzschild metric, we have of course neglected any small gravitational field which the scalar wave itself might produce.

II. GENERAL CONSIDERATIONS

In flat space-time the metric tensor in spherical coordinates is

$$g_{\mu\nu}=\mathrm{diag}[1,\ -1,\ -r^2,\ -r^2\sin^2\Theta],$$

and the scalar wave equation

$$\Box\Psi \equiv g^{\mu\nu}\Psi_{;\mu\nu}=0,$$

is separable.

In Schwarzschild space-time, the metric tensor is

$$g_{\mu\nu}=\mathrm{diag}\left[\left(1-\frac{r_s}{r}\right),\ -\left(1-\frac{r_s}{r}\right)^{-1},\ -r^2,\ -r^2\sin^2\Theta\right],$$

where r_s is related to the mass M by the relation $r_s=2M$, and the equation which determinates the scattering is a generalization of the flat space-wave equation.

If

$$\Psi=R(r)\,y(\Theta,\ \phi)\exp(-i\omega t),$$

then $Y(\Theta,\phi)$ is a spherical harmonic, and we have for $R(r)$, the radial wave equation

$$r(r-r_s)^2\frac{d^2R}{dr^2}+(r-r_s)(2r-r_s)\frac{dR}{dr}$$

$$+[k^2r^3-l(l+1)(r-r_s)]R=0. \tag{1}$$

where $k=\omega$.

We use now a new coordinate (the Regge—Wheeler coordinate r^*)

$$r^*=r+r_s\ln\left(\frac{r}{r_s}-1\right),$$

so that as

$$r \to +\infty\,(r_s),\quad r^* \to (+\infty)(-\infty).$$

In terms of r^*, the radial part of the wave equation reads

$$\frac{d^2R}{dr^{*2}}+\left\{k^2-l(l+1)/r^2+r_s/r^3\right.$$

$$\left.\times[r_s/r+l(l+1)-1]\right\}R(r^*)=0. \tag{2}$$

Equation (2) is similar to the one-dimensional Schrödinger equation with independent variable r^*. This equation has an effective potential

$$V_{\text{ef}}(r^*)=\left(1-\frac{r_s}{r}\right)\left(\frac{r_s}{r^3}+\frac{l(l+1)}{r^2}\right), \tag{3}$$

with r considered as a function of r^*.

Part of this effective potential $[l(l+1)/r^2]$, is the "centrifugal barrier," and part $[r_s/r]$, is due to the curvature of space-time.

This scalar wave potential is positive for all $r>2M$. It rises from 0 at $r=2M$ to a barrier summit, then, falls back to 0 at $r=\infty$.

We can also write Eq. (1) in the Schrödinger form, by means of the substitution

$$R(r)=\beta(r)\cdot g(r),$$

with

$$\beta(r)=k^{-1}/\sqrt{r(r-r_s)}.$$

We obtain for $g(r)$,

$$\frac{d^2g}{dr^2}+[k^2-V(r)]g=0, \tag{4}$$

where

$$V(r)=k^2\left[1-\frac{1}{(1-r_s/r)^2}\right]+\frac{l(l+1)}{r^2(1-r_s/r)}-\frac{r_s^2}{4r^4(1-r_s/r)^2}\,. \tag{5}$$

The function $V(r)$ can be interpreted also as an effective potential. As we have pointed out in Sec. I, its singular behavior near the horizon

$$V(r)=-\frac{\gamma}{(r-r_s)^2}+O\left(\frac{1}{r-r_s}\right),$$

where $\gamma=k^2r_s^2+\frac{1}{4}$, is responsible of the capture of waves by the black hole. Due to the fact that $\gamma>\frac{1}{4}$, the absorption will be present at all energies.[3]

III. HIGH ANGULAR MOMENTA–JWKB APPROXIMATION

For large values of l, waves of short wavelength $(kr_s\gg1)$ can be analyzed to good approximation by the quasiclassical method. Hence, the phase of the radial wave function $R(r^*)$ (Eq. 2), is given by the well-known JWKB expression,

$$\int_{r_0^*}^{r^*}[k^2-V_{\text{ef}}(r^*)]^{1/2}\,dr^*+\pi/4, \tag{6}$$

where r_0 is the classical turning point. This expression is not valid if r_0 approaches the maximum of the potential of Eq. (3).[4] That is, formula (6) holds for $(l+\frac{1}{2})=kb\gg kr_s$.

Due to the slow decrease of the effective potential, the wave is distorted even at large distances by the presence of a logarithmic term in the phase. This term is entirely produced by the Coulomb tail of the effective potential. The radial wave function can be written as

$$R_l(r)\sim(1/r)\sin[kr+kr_s\log vr+\delta_l(k)-l\pi/2]+O\left(\frac{1}{r^2}\right), \tag{7}$$

where v is a constant with inverse length dimensions.

We follow Matzner convention[1] in order to fix $\delta_l(k)$ unambiguously, that is, we take $v=2k$. Then, with the substitution

$$dr^*=dr(1-r_s/r)^{-1}, \tag{8}$$

we get from Eqs. (6) and (7),

$$\delta_l(k)=\lim_{r\to\infty}\left[\int_{r_0}^{r}\left([k^2-V_{\text{ef}}(r)]^{1/2}\,\frac{1}{1-(r_s/r)}-k\right)\times dr+\tfrac{1}{2}\pi(l+\tfrac{1}{2})-kr_0-kr_s\ln 2kr\right]. \tag{9}$$

This JWKB phase shift can be expressed in terms of a combination of complete and uncomplete elliptic integrals of first, second and third kinds. Due to the complex form of this expression we will proceed to calculate $\delta_l(k)$ in an approximate form.

For large impact parameters, as the gravitational field acting on the wave is weak, we will expand $\delta_l(k)$ in powers of r_s/b. By integration, one obtains, from Eq. (9),

$$\delta_l(k)=-kr_s\ln(l+\tfrac{1}{2})-\frac{kr_s}{2}\left[1+\frac{1}{(l+\frac{1}{2})^2}\right]+\frac{1}{3}\,\frac{(kr_s)^{3/2}}{(l+\frac{1}{2})^{1/2}}+\frac{15\pi}{32}\,\frac{(kr_s)^2}{(l+\frac{1}{2})}+O\left(\frac{1}{l+\frac{1}{2}}\right)^{3/2}. \tag{10}$$

The JWKB phase shift is connected with the classical function $\Theta(l)$ of the same angular momenta by[3]

$$\Theta(l)=2\,\frac{d}{dl}\,\delta_l^{\text{JWKB}}.$$

Then

$$\Theta(b)=-\frac{4M}{6}+\frac{M}{\pi^2b^3}\lambda^2-\frac{1}{3}\left(\frac{r_s}{b}\right)^{3/2}-\frac{15}{16}\,\pi\left(\frac{r_s}{b}\right)^2.$$

The second term depends upon the energy of incident plane wave. The short wavelength limit gives Einstein deflection, $\Theta=-4M/b$, for large impact parameters.

Making use of the eikonal approximation,[5] we can find the scattering amplitude $f(\theta)$, for small angles, as

$$f(\theta\neq0)=-ik\int_0^\infty db\;b\;J_0(kb\theta)\exp[2i\delta_l^{\text{JWKB}}(k,b)]$$

Using the δ_l^{JWKB}, [Eq. (10)] to order kr_s, we find

$$f(\theta\neq0)=\frac{i}{2k}\exp\left\{(ikr_s)\left[(2\log\theta/2-1)-2\arg\Gamma(ikr_s)\right]\left[1+\frac{4}{\theta^2}(ikr_s)\right]\right\}.$$

The differential cross section is

$$\frac{d\sigma}{d\Omega}=\frac{16M^2}{\theta^4}+\frac{\lambda^2}{16\pi^2}\,. \tag{11}$$

This expression gives the Rutherford law for small angle scattering cross section more wavelength-dependent correction.

All results up to here are valid for $b\gg r_s$ and small deflection angles. For $b\sim r_s$, the wave absorption process is important and must be taken into account. Now, we will calculate the partial wave absorption coefficient. We can find in a simple calculation the JWKB absorption

coefficient for waves with impact parameter such that k^2 is near the top of the effective potential. One sees from expression (3) that the maximum of V_{ef} occurs at

$$r_{max}' = \tfrac{3}{2} r_s \left[1 - \frac{1}{q(l+\frac{1}{2})^2}\right] + O\left[\frac{1}{(l+1/2)^4}\right].$$

For r sufficiently close to r_{max}', we can write

$$k^2 - V_{ef} \approx k^2 - V_{ef}(r_{max}') - \tfrac{1}{2} H_0 (r - r_{max}')^2,$$

where

$$k^2 - V_{ef}(r_{max}') = k^2 - \tfrac{8}{81}\left[1 + \tfrac{3}{2}(l+\tfrac{1}{2})^2\right] + O[1/(l+\tfrac{1}{2})^4],$$

and

$$H_0 = \left(\frac{2}{3}\right)^6 \frac{7}{9 r_s^4}\left[1 - \tfrac{9}{14}(l+\tfrac{1}{2})^2\right] + O\left[\frac{1}{(l+1/2)^2}\right].$$

The quasiclassical approximation yields the following expression for the absorption coefficient[3]

$$D_l^{JWKB} = 1/(1 + \exp(-2\pi\epsilon)),$$

where

$$\epsilon = (1/2\sqrt{H_0})\,[k^2 - V_{ef}(r_{max}')].$$

We obtain

$$D_l^{JWKB} = \frac{1}{1 + \exp[2\pi(l+\frac{1}{2})][1 - (27k^2 r_s^2/4(l+1/2)^2)]} + O\left[\frac{1}{(l+1/2)^3}\right]. \tag{12}$$

The total capture cross section is given by

$$\sigma_{capt} = \frac{\pi}{k^2} \sum_{l=0}^{\infty} (2l+1) D_l. \tag{13}$$

For large values of kr_s, the partial wave series can be approximated by an integral because the main contribution comes from the higher angular momenta. Correspondingly, we see from Eq. (13) that the capture cross section is

$$\sigma_{capt} = 2\pi \int_0^{\infty} b\, D(b)\, db. \tag{14}$$

If we take for the absorption coefficient the JWKB expression given by Eq. (12), the integral in Eq. (14) can be solved exactly.[6] One obtains

$$\sigma_{capt} = \tfrac{27}{4}\pi r_s^2 + \lambda^2/24\pi,$$

in agreement with the geometrical optical result for $kr_s \to \infty$.[7] The absorption cross section is increased by the λ-dependent corrections, because the waves can "tunnel" through the potential barrier.

IV. SCATTERING FOR LOW ANGULAR MOMENTA

We will study now the radial wave equation (4) for the low l case, $l \ll kr_s$. As we are interested in the short wavelength solutions, this case corresponds to $b \ll r_s$. In this case, the capture coefficient is expected to be large.

When $r \to r_s$, we see by expression (5), that

$$V(r) \to -\left(k^2 + \frac{1}{4r_s^2}\right) \frac{1}{(1 - r_s/r)^2} + O\left(\frac{1}{r - r_s}\right).$$

For $b \ll r_s$, the wave "will see" mainly this internal part of the potential. By this reason, we consider as the approximate potential

$$V_0(r) = -k^2 - \frac{k^2 + 1/4r_s^2}{(1 - r_s/r)^2} + \frac{1}{4r_s^2}\left[1 + \frac{2}{(r/r_s) - 1}\right].$$

This expression taken for $V_0(r)$ reproduces the behavior of the exact potential for $r \to r_s$ and also gives the correct behavior for $r \to \infty$.

The radial wave equation (4) for the potential V_0 can be solved exactly, and the effective interaction $V(r) - V_0(r)$ will be treated as a perturbation. This approximation is expected to be good for

$$l \ll k\, r_s \gg 1.$$

Introducing the variable

$$z = r/r_s - 1,$$

we solve for the potential V_0,

$$\frac{d^2 g_0}{dz^2} + \left[\frac{k^2 r_s^2 + 1/4}{z^2} + 2\,\frac{k^2 r_s^2}{z} + k^2 r_s^2\right] g_0(z) = 0, \tag{15}$$

which corresponds to a purely attractive potential.

The linearly independent solutions of the radial Eq. (15) can be written as

$$g_0^{(-)}(z) = \sqrt{z}\, \exp[-ikr_s(z + \log z)] \times F(\tfrac{1}{2},\ 1 - 2ikr_s,\ 2ikr_s z) \tag{16}$$

$$g_0^{(+)}(z) = \sqrt{z}\, \exp[ikr_s(z + \log z)] \times \Psi(\tfrac{1}{2},\ 1 + 2ikr_s,\ -2ikr_s z), \tag{17}$$

where F and Ψ are the confluent hypergeometric functions of the first and second kinds, respectively.[8]

Both solutions are well-behaved at $z = 0$ ($r = r_s$). This is so, because the effective potential is singular and attractive at $z = 0$. This means that from the point of view of regularity, both functions are physically acceptable. Nevertheless, the function $g_0^{(-)}$ corresponds to waves going into the black hole, while $g_0^{(+)}$ describes outgoing waves at the Schwarzschild radius. By this reason, we choose $g_0^{(-)}$[1] as the physical solution. We are interested in the asymptotic behavior for $z \to \infty$. Making use of the asymptotic development of (16) for large z,[8] we obtain

$$g_0^{(-)}(r)_{r\to\infty} = \exp[i(kr_s + \pi/4)]\, \exp[(ikr_s - \tfrac{1}{2}) \times \log 2kr_s] \frac{\Gamma(1 - 2ikr_s)}{(-1)^l \Gamma((1/2) - 2ikr_s)} \cdot \Big\{(-1)^l \times \exp[-i(kr + kr_s \log 2kr)] - \exp[2i\delta_l^{(0)}] \exp[i(kr + kr_s \log 2kr)]\Big\},$$

with

$$\exp(2i\delta_l^{(0)}) = (-1)^{l+1}\, \frac{\exp[i(\delta_\phi - 2kr_s - \pi/2)]}{\sqrt{(1/2)(1 + \exp(4\pi kr_s))}},$$

where

$$\delta_\phi = \arg\Gamma(\tfrac{1}{2} - 2ikr_s).$$

Finally,

$$\mathrm{Re}\delta_l^{(0)} = \frac{\delta_c}{2} + (\pi/2)(l+\tfrac{3}{2}) - kr_s, \tag{18}$$

$$\mathrm{Im}\delta_l^{(0)} = \tfrac{1}{4}\log\left(\frac{1+\exp(4\pi kr_s)}{2}\right). \tag{19}$$

We see that the partial amplitude $\exp(2i\delta_l^{(0)})$ results in a complex quantity whose modulus is less than unity. The physical meaning is the presence of capture processes. The fraction of the wave that is captured is given by the partial wave absorption coefficient

$$P_l^{(0)} = 1 - |\exp(2i\delta_l^{(0)})|^2.$$

From (19),

$$P_l^{(0)} = \frac{1-\exp(-4\pi kr_s)}{1+\exp(-4\pi kr_s)}.$$

For $kr_s \gg 1$,

$$P_l^{(0)} = 1 - 2\exp(-4\pi kr_s).$$

The absorption coefficient increases with kr_s, this is with the energy, as expected.

In order to improve the previous results for $\delta_l(k)$ and $P_l(k)$, we will consider the perturbation $V_0 - V$. The exact radial equation (4) can be written in the form

$$\frac{d^2g}{dz^2} + [k^2r_s^2 - r_s^2V_0(z) - r_s^2V_1(z)]g(z) = 0, \tag{20}$$

with

$$V_1(z) = \frac{1}{z(z+1)r_s^2}\left[(l+\tfrac{1}{2})^2 + \frac{1}{4(z+1)}\right].$$

The Green's function corresponding to Eq. (20), which satisfies

$$\left[\frac{d^2}{dz^2} + k^2r_s^2 - r_s^2V_0(z)\right]G_0(z,z') = \delta(z-z'),$$

is given by

$$G_0(z,z') = \mathbb{C}\, g_0^{(-)}(z_<)\circ g_0^{(+)}(z_>),$$

with $g_0^{(-)}$ and $g_0^{(+)}$ given by (16) and (17), and where the constant $\mathbb{C}$ is equal to

$$\mathbb{C} = \Gamma(\tfrac{1}{2} - 2ikr_s)/\Gamma(1-2ikr_s).$$

Then the differential equation (20) with the chosen boundary conditions is equivalent to the following integral equation,

$$g(z) = g_0^{(-)}(z) + r_s^2\int_0^\infty G_0(z,z')\,V_1(z')\,g(z')\,dz'.$$

We now make the first approximation of the Neuman–Liouville series (DWBA) and, for large distances, we obtain

$$g(z)_{z\to\infty} \sim g_0^{(-)}(z)_{z\to\infty} + r_s^2\mathbb{C}\; g_0^{(+)}(z)_{z\to\infty}\circ \mathbb{I}, \tag{21}$$

$$\mathbb{I} = \int_0^\infty [g_0^{(-)}(z')]^2\,V_1(z')\,dz'.$$

We obtain (see Appendix)

$$\mathbb{I} = \Gamma(1-2ikr_s)(2kr_s)^{2ikr_s}\exp(-\pi kr_s)\circ\left[\tfrac{1}{4}\exp(i\pi/2) + (l+\tfrac{1}{2})^2\exp(i\pi/4)\left(\frac{\pi}{kr_s}\right)^{1/2} + O\left(\frac{1}{kr_s}\right)^{3/2}\right]. \tag{22}$$

From (21) and (22), one gets

$$\mathrm{Re}\delta_l(k) = \frac{\delta_c}{2} - kr_s + \frac{\pi}{2}(l+\tfrac{3}{2}) + \frac{1}{16}\left(\frac{\pi}{kr_s}\right)^{1/2} + O\left(\frac{1}{kr_s}\right)^{3/2},$$

$$\mathrm{Im}\delta_l(k) = \tfrac{1}{4}\log\left[\frac{1+\exp(4\pi kr_s)}{2}\right] - \frac{1}{16}\left(\frac{\pi}{kr_s}\right)^{1/2} - \frac{\pi}{kr_s}\,\frac{(l+1/2)^2}{2\sqrt{2}} + O\left(\frac{1}{kr_s}\right)^{3/2}.$$

Then, the absorption coefficient in the DWBA is given by

$$P_l^{\mathrm{DWBA}} = \frac{1-\exp(-4\pi kr_s)}{1+\exp(-4\pi kr_s)} - \frac{2\exp(-4\pi kr_s)}{1+\exp(-4\pi kr_s)}\cdot\left\{\exp\left[\frac{1}{4}\left(\frac{\pi}{kr_s}\right)^{1/2} + \sqrt{2}\left(\frac{\pi}{kr_s}\right)(l+\tfrac{1}{2})^2 + O(kr_s)^{-3/2}\right]\right\}.$$

The first term, $P_l^{(0)}$, corresponds to the approximative potential V_0, while the second term gives the first-order contribution due to the potential V_1. As is seen from Eq. (23), this second term is a small correction to the $P_l^{(0)}$ because $kr_s \gg 1$.

The sign of the contribution of V_1 is related to its repulsive character. As is seen from Eq. (23), P_l^{DWBA} tends to one for kr_s going to infinity. However, the JWKB absorption coefficient (Eq. 12) remains always much less than one in its range of validity. In other words, the lower angular momenta partial waves are more absorbed than the higher ones.

V. CONCLUDING REMARKS

We have applied and adapted to the scattering of waves from a Schwarzschild black hole, approximation methods appropriate for small wavelengths. In this way, we have obtained explicitly the leading behavior on the wavelength of several scalar scattering parameters.

The methods used here can be applied to the neutrino equation, as the electromagnetic and gravitational wave scattering, in the Schwarzschild geometry.

The formalism exhibited here can be generalized to the wave scattering and the quantum processes of particle production in the Kerr–Newmann geometry.

APPENDIX

With the aid of formulas (3),

$$\int_0^\infty \exp(-\lambda z)\,z^{\gamma-1}[F(\alpha,\gamma,kz)]^2\,dz = \Gamma(\gamma)\,\lambda^{2\alpha-\gamma}(\lambda-k)^{-2\alpha}\,F\left(\alpha,\alpha,\gamma;\frac{k^2}{(\lambda-k)^2}\right),$$

and

$$F(\alpha,\beta,\gamma;z) = (1-z)^{\gamma-\alpha-\beta}\,F(\gamma-\alpha,\ \gamma-\beta,\ \gamma;z),$$

where $F(\alpha,\beta,\gamma,z)$ is the Gauss' hypergeometric function, we have reduced the integral

$$\mathbb{I}_1 = (l+1/2)^2 \int_0^\infty \exp(-2ikr_s z)\ \frac{z^{-2ikr_s}}{(z+1)}$$

$$\times [F(\tfrac{1}{2},\ 1-2ikr_s,\ 2ikr_s z)]^2\ dz\,,$$

corresponding to the first term of the perturbation V_1, to the following expression:

$$\mathbb{I}_1 = (l+1/2)^2\,\Gamma(1-2ikr_s)$$

$$\times \int_0^\infty dz\,(z-2ikr_s)^{-2ikr_s} z^{4ikr_s-1}\, e^{-z}$$

$$\cdot F(\tfrac{1}{2}-2ikr_s,\ \tfrac{1}{2}-2ikr_s,\ 1-2ikr_s,\ -4k^2r_s^2/z^2).$$

The asymptotic expansion for $kr_s \gg 1$, of the integral $\mathbb{I}_1$ can be calculated by the stationary phase method,[9]

$$\mathbb{I}_1 = (l+1/2)^2\,\Gamma(1-2ikr_s)(2kr_s)^{2ikr_s}$$

$$\times \exp(-\pi kr_s + i\pi/4)\cdot \left(\frac{\pi}{kr_s}\right)^{1/2} + O\left(\frac{1}{kr_s}\right)^{3/2}.$$

The integral corresponding to the second term of V_1 follows immediately from the previous results.

ACKNOWLEDGMENTS

We thank H. J. de Vega for valuable suggestions and helpful discussions. We thank J. R. Albano and S. Bonazzola for useful remarks and valuable encouragement. I would like to express my appreciation to Instituto de Astronomía y Física del Espacio, where part of this work was realized.

*On leave on absence from Institute of Astronomy and Spacial Physics (CONICET, Buenos Aires, Argentina).

[1]R. A. Matzner, J. Math. Phys. 9, 163 (1968).
[2]S. Persides, J. Math. Phys. 14, 1017 (1973); 15, 885 (1974).
[3]L. D. Landau and E. M. Lifschitz, *Quantum Mechanics* (Pergamon, London, 1965).
[4]W. Ford and J. A. Wheeler, Ann. Phys. (NY) 7, 239 (1959).
[5]R. G. Newton, *Scattering Theory of Waves and Particles* (McGraw-Hill, New York, 1966).
[6]I. S. Gradsteyn and I. M. Ryzhik, *Table of Integrals, Series and Products* (Academic, New York and London, 1965).
[7]C. W. Misner, K. S. Thorne, and J. A. Wheeler, *Gravitation* (Freeman, San Francisco, 1973).
[8]N. N. Lebedev, *Special Functions and Their Applications* (Prentice-Hall, Englewood Cliffs, New Jersey, 1965).
[9]A. Erdélyi, *Asymptotic Expansions* (Dover, New York, 1956).

A geometric proof of no-interaction theorems*

Aloysius F. Kracklauer†

Department of Physics, University of Houston, Houston, Texas 77004
(Received 30 December 1974)

No-interaction theorems are proven, using the methods of modern differential geometry, and an example of a Hamiltonian yielding relativistic canonical equations of motion with interaction is presented.

1. INTRODUCTION

It is the purpose of this note to present a proof of "no-interaction" theorems using the language of modern differential geometry[1] and Cartan's principle[2] of dynamics. The use of these tools reveals certain facets of the structure of these theorems which are not otherwise evident.

The conclusion of a no-interaction theorem is that a relativistic canonical formulation of dynamics can only describe particles between which there is "no interaction." Of course, this conclusion depends on the conditions contained in the hypothesis of the theorem, of which there are two versions. The original version has been proven sequentially by stronger methods, first for two,[3] then three,[4] and finally N particles by at least three methods,[5-7] one of which (Ref. 7) uses modern differential geometry, but not Cartan's principle. This version is characterized by the assumption that the dynamics of the system is governed by a scheme parameterized by a single parameter—time. The second version,[8] which has been proved heretofore by only one method, is characterized by the assumption that the dynamics is governed by an N parameter scheme.

Below, the no-interaction result is obtained in a new way by adding restrictions to Cartan's principle that are equivalent to the conditions in the hypotheses of the no-interaction theorems. From this unique vantage point these theorems are reexamined.

2. A GEOMETRIC PROOF

Let (M_N, Ω) be a symplectic manifold,[1] where M_N is the Cartesian product of N phase spaces with a fundamental two-form

$$\Omega = d\omega. \tag{1}$$

ω is expressed in terms of the canonical momenta and coordinates as

$$\omega = \sum_i^N (p^\mu dx_\mu)_i, \tag{2}$$

(summation convention implied). Recall[2] that a one-form on $M_N \times I$, for example ω':

$$\omega' = \omega - H d\tau \tag{3}$$

determines a vector field D on $(M_N \times I)$ of the form

$$D = \sum_i^N \left(V^\mu \frac{\partial}{\partial p^\mu} + F_\mu \frac{\partial}{\partial x_\mu} \right)_i + \frac{\partial}{\partial \tau} \tag{4}$$

via the stipulation that the exterior product of the pair is zero (this is the formal statement of Cartan's principle); i.e.,

$$D \lrcorner d\omega' = 0, \tag{5}$$

such that V and F satisfy Hamilton's canonical equations:

$$V = \frac{\partial H}{\partial p}, \qquad F = -\frac{\partial H}{\partial x}. \tag{6}$$

Theorem (no interaction): Let (M_N, Ω) be a symplectic manifold, D a vector field on $M_N \times I$, and H a scalar function. Moreover, let I be identified with a one-dimensional subspace from each configuration space; for example, let

$$\tau = a x_i^1 = b x_j^2 = \cdots = c x_k^N. \tag{7}$$

Then,

$$D \lrcorner d\omega' = 0 \tag{8}$$

implies that

$$F_1 = F_2 = \cdots = F_N = 0; \tag{9}$$

i.e., all forces are zero, there is no interaction.

Proof: The identification of I with a one-dimensional subspace of the configuration space of a particle implies that the generator of translations along I, namely

$$\frac{\partial}{\partial \tau}, \tag{10}$$

is equal to the generators of translations along the one-dimensional subspace in the configuration space, for example,

$$a \frac{\partial}{\partial x_i^1}, \tag{11}$$

where it has been assumed for simplicity, without loss of generality, that the identification has been made the ith axis in the coordinate frame chosen to describe the configuration space of interest. When this identification is made for each configuration space in M_N, the following equality holds:

$$\frac{\partial}{\partial \tau} = a \frac{\partial}{\partial x_i^1} = b \frac{\partial}{\partial x_j^2} = \cdots = c \frac{\partial}{\partial x_k^N}. \tag{12}$$

If (12) is put into (8), then the result (9) follows at once. ∎

Theorem (second version): The fundamental assumption of this version is that the dynamics of a system of particles is governed by an N parameter scheme. This

assumption is tantamount to redefining Eq. (4) as

$$D=\sum_i^N\left[\left(V^\mu\frac{\partial}{\partial p^\mu}+F_\mu\frac{\partial}{\partial x_\mu}\right)_i+\frac{\partial}{\partial\tau_i}\right], \tag{13}$$

and Eq. (3) as

$$\omega'=\omega-\sum_i^N H_i\,d\tau_i. \tag{14}$$

Proof: As above, (8) is computed to obtain

$$\{H_i, H_j\}\equiv 0, \quad \forall\, i\neq j. \tag{15}$$

where the brackets are those of Poisson. The independence of world lines follows from the independence of the Hamiltonians. ∎

3. DISCUSSION AND CONCLUSIONS

The equivalence of the first version as presented herein with the previous presentations follows from the demands made of the Hamiltonian in those versions. First, and most naturally, the Hamiltonian is to specify the dynamics through the Lie bracket relationships

$$[x, H]=V, \quad [p, H]=F. \tag{16}$$

Beyond this, however, the very same Hamiltonian function is also to be the generator of time translations which together with the generators of the spatial translations form a Lie group with the Poincaré group structure constants. Imposing this group structure is a common way of "relativising" the space of interest. Demanding that one function yield both sorts of generators at once establishes the identification used above of translations along τ (the parameter of the transformation specifying the dynamics) and x_i^n (a configuration parameter, most likely t if M_n is a Minkowski space).

The absence of reference to the Minkowski structure of the configuration spaces in the above proof shows that the essential feature of no-interaction theorems is not to be found in special relativity. The above theorem holds even if the spaces M_n are not Minkowski spaces; the only essential needed to obtain the no-interaction result is the identification of the parameter of the transformation generating the dynamics with any of the parameters of the configuration spaces. In the nonrelativistic case, if the time "t" were identified with either x, y, or z (or linear combinations), the no-interaction result would follow, a fact not heretofore widely publicized.

The motivation for the identification of the dynamical parameter with a configuration parameter in the first place comes from the desire to create relativistic quantum theories. In nonrelativistic quantum mechanics time is a scalar parameter while configuration variables are operators. This disparity is at odds with special relativity within which time and space parameters are of equal status. Thus, an effort has been made to find a relativistic formulation of mechanics in which the configuration variable "t" (x_4 in Minkowski space) can serve as the independent parameter of the transformation specifying the dynamics and thereby be compatible with quantum theory as presently formulated. It is these efforts which are frustrated by the first version of no-interaction theorems.

The second version is an exploratory attempt to find a structure that will accommodate interaction and be consistent with relativity. As the structure hypothesized in this version is not compatible with quantum theory, it does not appear to have any significance for the construction of relativistic quantum theory; moreover, it also does not accommodate interaction.

If the identification of time and the parameter which governs the dynamics is not made, as it is in the original version of no-interaction theorems, then Cartan's formulation of dynamics (or any equivalent) accommodates the construction of canonical relativistic theories at will; it is only necessary to find a suitable single parameter Lorentz invariant scalar function H. What is not so easily done, however, is to show that theories so constructed do in fact describe interactions employed by nature. This can be done only by showing that a particular choice of H leads to calculated trajectories that have observable correspondents. For an example of a Hamiltonian that may describe the electromagnetic interaction, consider the following: Let m_i be the rest mass of the ith particle and let $\mathbf{x}_i$ and $\dot{\mathbf{x}}_i$ be defined as follows:

$$\mathbf{x}_i\triangleq x_i,\ y_i,\ z_i,\ ict_i;\quad \mathbf{x}_i\triangleq d\mathbf{x}_i/d\tau. \tag{17}$$

If now the Lagrangian L, where

$$L(\mathbf{x}_i(\tau),\dot{\mathbf{x}}_i(\tau))=\sum_i^N\frac{m_i\dot{\mathbf{x}}\cdot\dot{\mathbf{x}}_i}{2}-2\sum_{j\neq i}^N e_ie_j \times\int_{-\infty}^{\tau}\dot{\mathbf{x}}_i\cdot\dot{\mathbf{x}}_j\,\delta((\mathbf{x}_i(\tau)-\mathbf{x}_j(\lambda))^2)\,d\lambda, \tag{18}$$

is posited (dot products are with respect to the Lorentz metric), then by employing the well-known definition of canonical momentum,

$$p_i\triangleq\frac{\partial L}{\partial\dot{\mathbf{x}}_i}, \tag{19}$$

$$H(\mathbf{x}_i(\tau),p_i(\tau))=\sum_i^N\left(p_i-2\sum_{j\neq i}^N e_ie_j \times\int_{-\infty}^{\tau}\dot{\mathbf{x}}_j\,\delta((\mathbf{x}_i(\tau)-\mathbf{x}_j(\lambda))^2)\,d\lambda\right)^2\Big/2m_i, \tag{20}$$

is deduced. This Hamiltonian leads to equations of motion which are differential-delay equations of motion coupled together by two and only two interactions, each derived from a Lienard–Wiechert potential. Although Cauchy-type initial data is insufficient to determine a particular solution to these equations, they can be integrated numerically given the orbits between the past and future of a light cone centered at an arbitrary point as initial data. The results of such a study will be reported elsewhere; the point here is only that Cartan's principle does accommodate canonical relativistic dynamics with interaction if the effort to give time a role distinct from space is abandoned.

*The results contained herein constitute part of a dissertation submitted to the University of Houston.

†Present address: 2344 Antigua Ct., Reston, Va. 22091.

[1]R. Abraham, *Foundations of Mechanics* (Benjamin, New York, York, 1967).
[2]W. Slebodzinski, Ann. Soc. Math. Poland Ser. **1**, 1 (1970).
[3]D.G. Currie, T.F. Jordan, and E.C.G. Sudarshan, Rev. Mod. Phys. **35**, 350 (1963).
[4]J.T. Cannon and T.F. Jordan, J. Math. Phys. **5**, 299 (1964).
[5]H. Leutwyler, Nuovo Cimento **37**, 556 (1965).
[6]R.N. Hill, J. Math. Phys. **8**, 1956 (1967).
[7]L. Bel, Ann. Inst. H. Poincaré A **14**, 189 (1971).
[8]P. Droz-Vincent, Nuovo Cimento B **12**, 1 (1972).

The Kirkwood–Salsburg equation and phase transitions

W. Klein

Institut für Theoretische Physik, Universität zu Köln, 5 Köln 41, Germany
(Received 21 July 1975)

A perturbation expansion solution of the Kirkwood–Salsburg equation is used to investigate the hard sphere phase transition. A necessary condition for the phase transition identical in form to one obtained by Kirkwood is obtained. However, it is shown that this condition is not sufficient and that other conditions must also be satisfied.

I. INTRODUCTION

In a previous paper[1] a perturbation expansion of the solution to the Kirkwood—Salsburg equation was introduced. It was shown in that paper that each term in the expansion for a given distribution function was obtained by solving an integral equation, the kernel of which remains the same to each order. It has been proved that this expansion converges only for small z; it is possible, however, that—due to the alternating signs of the terms in the series—the radius of convergence is much larger than the proven one. To test this possibility, the properties of the perturbation solution for hard spheres are investigated in the region of $Z = Z_c$ where multiple solutions seem to be possible. These solutions are characterized by a single particle distribution function which is not constant but oscillates as a function of X_1 with a fixed period.

A necessary condition for such a solution is derived. The condition is identical in form to the condition Kirkwood[2] derived for the instability of the solutions to the Born—Green hierarchy. We also show that this condition cannot be considered to indicate a first order phase transition but is at best an indication of an instability.

II. PERTURBATION EXPANSION

In this paper we consider only the hard sphere interaction, i.e.,

$$\phi(r) = 0, \quad r > \sigma,$$
$$\phi(r) = \infty, \quad r \leq \sigma,$$

although the results can be extended to all symmetric positive pair potentials such that

$$\phi(r) = 0, \quad r > A.$$

We will consider the K—S equation

$$\rho_1(x_1) = z\left(1 + \sum_{n=1}^{\infty} \frac{1}{n!} \int \rho_n(x_2 \cdots x_{n+1}) \prod_{j=2}^{n+1} f_{ij}\, dx_j\right),$$

$$\rho_N(x_1 \cdots x_N) = z \prod_{j=2}^{N} (1 + f_{1j}) \left(\rho_{N-1}(x_a \cdots x_N) + \sum_{n=1}^{\infty} \frac{1}{n!} \int \rho_{N+n-1}(x_2 \cdots x_{N+n}) \prod_{j=N+1}^{N+n} f_{1j}\, dx_j\right),$$

$$f_{ij} = \exp[-\beta\phi(|x_{ij}|)] - 1, \qquad (\text{II}.1)$$

which can be written concisely as

$$\rho = z + zK\rho,$$

$$\rho = \begin{pmatrix} \rho_1(x_1) \\ \vdots \\ \rho_N(x_1 \cdots x_N) \\ \vdots \end{pmatrix}.$$

K is a bounded linear operator.[3] K can be broken up into a perturbation K' [4] which we argue has a small effect[4] and an unperturbed operator K_1. K' is defined as

$$K'\rho = \begin{pmatrix} 0 \\ (1 + f_{12})\rho_1(x_2) \\ \vdots \\ \prod_{j=2}^{N} (1 + f_{1j})\rho_{N-1}(x_2 \cdots x_N) \\ \vdots \end{pmatrix}$$

and

$$K = K' \neq K_1.$$

If we consider the equation[4]

$$\rho = z\alpha + z(K_1 + \epsilon K')\rho, \quad \alpha = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix},$$

expand ρ in a power series in ϵ, equate powers of ϵ, and solve, we obtain for $\epsilon = 1$

$$\rho = \sum_{n=0}^{\infty} [(1 - zK_1)^{-1} zK']^n (I - zK_1)^{-1} \alpha.$$

One can also write an equation for the first nonzero contribution to each of the distribution[1] functions

$$\rho_1(x_1) = z + z \int \rho_1(x_2) f_{12}\, dx_2$$
$$\vdots$$
$$\rho_N(x_1 \cdots x_N) = z \prod_{j=2}^{N} (1 + f_{1j}) [\rho_{N-1}(x_2 \cdots x_N) + \int \rho_N(x_{N+1}, x_2 \cdots x_N) f_{1,N+1}\, dx_{N+1}$$
$$\vdots \qquad (\text{II}.2)$$

It is important to note, as was pointed out in Ref. 1, p. 1052, that solving the K—S equation by means of this method involves the solution only of (II.2) with changing inhomogeneous terms. Therefore, the properties of the solution(s) of the (K—S) hierarchy which depend only on the kernels in (II.2) will be the same for

each order. Clearly the existence of multiple solutions is such a property. We will investigate therefore the solutions of the above first order hierarchy.

III. UNIQUENESS OF SOLUTION

We will first investigate the uniqueness of the solutions of (II.2) for the hard sphere potential. We will prove that for values of z for which

$$1 - z\hat{f}(|\mathbf{K}|) \neq 0$$

$$\hat{f}(|\mathbf{K}|) = \int \exp(-i\mathbf{K}\cdot\mathbf{x}) f(|\mathbf{x}|)\, d\mathbf{x} \qquad \text{(III.1)}$$

for every K real the solution of (II.2) is unique. Clearly, if each equation in (II.2) has a unique solution than the total hierarchy (II.2) has a unique solution.

It is trivial to prove, using the convolution theorem for the Fourier transform, that under the assumed condition the first equation of (II.2) has a unique solution. We now prove that the solution of

$$\rho_N(x_1 \cdots x_N) = z \prod_{j=2}^{N} (1 + f_{1j}) [\rho_{N-1}(x_2 \cdots x_N) + \int \rho_N(x_{N+1}, x_2, \ldots, x_N) f_{1,N+1}\, dx_{N+1}] \qquad \text{(III.2)}$$

is also unique.

Proof: Since the operator is linear, we must prove that the only solution to

$$\rho_N(x_1 \cdots x_N) = z \prod_{j=2}^{N} (1 + f_{1j}) \int \rho_N(x_{N+1}, x_2, \ldots, x_N) \times f_{1,N+1}\, dx_{N+1} \qquad \text{(III.3)}$$

is $\rho_N = 0$. We assume there exists a nonzero solution $\rho_0(x_1 \cdots x_N)$ and define $\rho_N(x_1 \cdots x_N)$ by

$$\rho_N(x_1, \ldots, x_N) = z \int \rho_0(x_{N+1}, x_2 \cdots x_N) f_{1,N+1}\, dx_{N+1},$$

which can be written as

$$\rho_N(x_1 \cdots x_N) = \rho_0(x_1 \cdots x_N) + \rho_I(x_2 \cdots x_N). \qquad \text{(III.4)}$$

$\rho_I(x_1 \cdots x_N)$ is nonzero only if

$$|x_1 - x_{1j}| < \sigma \quad \text{for some } j \in \{2 \cdots N\}.$$

Using definition (III.4) and taking the Fourier transform gives

$$[1 - z\hat{f}(|\mathbf{K}|)]\hat{\rho}_0(\mathbf{K}, x_2 \cdots x_N) = -\rho_I(\mathbf{K}, x_2 \cdots x_N). \qquad \text{(III.5)}$$

As $[1 - z\hat{f}(|\mathbf{K}|)] \neq 0$ if, for all K, $\hat{\rho}_I(\mathbf{K}, x_2 \cdots x_N) = 0$, then for all K

$$\hat{\rho}_0(\mathbf{K}, x_2 \cdots x_N) = 0 \Rightarrow \rho_0(x_1, \ldots, x_N) = 0.$$

Therefore we assume $\hat{\rho}_I(\mathbf{K}, x_2 \cdots x_N)$ is not identically zero. We now multiply each side of (III.5) by $\rho_0^*(K, x_2 \cdots x_N)$ (* denotes complex conjugate) and integrate with respect to K

$$\int (1 - z\hat{f}(K)) |\hat{\rho}_0(\mathbf{K}, x_2 \cdots x_N)|^2 d\mathbf{K} = -\int \hat{\rho}_I(\mathbf{K}, x_2 \cdots x_N) \hat{\rho}_0(\mathbf{K}, x_2 \cdots x_N)\, d\mathbf{K}. \qquad \text{(III.6)}$$

The left-hand side of the above is positive definite, the right-hand side is identically zero as can be seen from the fact that

$$\int \rho_0^*(\mathbf{x}_1 \cdots x_N) \rho_I(\mathbf{x}_1 \cdots x_N) \exp(-i\mathbf{K}'\cdot\mathbf{x}_1)\, d\mathbf{x}_1 = 0$$

for all $\mathbf{K}'$,

$$= \int \rho_0^*(\mathbf{K} - \mathbf{K}', x_2 \cdots x_N) \rho_I(\mathbf{K}, x_2 \cdots x_N)\, d\mathbf{K},$$

setting $\mathbf{K}' = 0$. We have a contradiction and the desired result is proven.

We have therefore for such z that

$$1 - z\hat{f}(|\mathbf{K}|) \neq 0$$

for any $\mathbf{K}$ a unique solution to (II.2). This implies that the solution, via this expansion, of the exact K—S equation (II.1) cannot have a solid structure, i.e., the distribution functions are functions only of the distances between particles. This ceases to be the case when

$$1 - z\hat{f}(|\mathbf{K}|) = 0$$

for some $\mathbf{K}$. For such z the first equation in (II.2) has the solution

$$\rho_1(\mathbf{x}_1) = \frac{z}{1 - z\hat{f}(0)} + \sum_{\alpha} (A \exp(i\mathbf{K}_\alpha \cdot \mathbf{x}_1) + B \exp(-i\mathbf{K}_\alpha \cdot \mathbf{x}_1))$$

where

$$1 - z\hat{f}(|\mathbf{K}|) = 0.$$

For hard spheres

$$\hat{f}(|\mathbf{K}|) = (4\pi/|\mathbf{K}|^2)[\cos|\mathbf{K}| - (\sin|\mathbf{K}|)/|\mathbf{K}|],$$

and we obtain Kirkwood's[2] criterion with z playing the role of λ. We now show, however, that the solutions which correspond to the minimum value of $z = z_0$ for which

$$1 - z_0\hat{f}(|\mathbf{K}|) = 0$$

for some K cannot be those of a stable solid, and hence Kirkwood's criterion does not indicate the onset of the phase transition.

It should be pointed out that Kirkwood stated this as a possibility.

IV. ADDITIONAL CONDITIONS

To be stable, a solid must not be invariant under arbitrary rotations. This means that the K—S equation should have at least one solution (for given boundary conditions) which is not invariant under arbitrary rotations. In this section we show that the solutions corresponding to $z = z_0$ do not have this property. This leads to the introduction of additional criterion which must be satisfied along with the Kirkwood criterion.

Consider the equation

$$\rho_2(\mathbf{x}_1, \mathbf{x}_2) = z(1 + f_{12})[\rho_2(x_2) + \int \rho_2(\mathbf{x}_3, \mathbf{x}_2) f_{12}\, d\mathbf{x}_3].$$

If we perform an arbitrary rotation Δ of $\mathbf{x}_1$ about $\mathbf{x}_2$, we find

$$\rho_2(\mathbf{x}_1 + \boldsymbol{\Delta}, \mathbf{x}_2) = z(1 + f_{12})[\rho_1(x_2) + \int \rho_2(\mathbf{x}_3 + \boldsymbol{\Delta}, \mathbf{x}_2) f_{13}\, d\mathbf{x}_3].$$

Subtracting the two equations gives

$$\rho_2(\mathbf{x}_1, \mathbf{x}_2) - \rho_2(\mathbf{x}_1 + \boldsymbol{\Delta}, \mathbf{x}_2) = z(1 + f_{12}) \int (\rho_2(\mathbf{x}_3, \mathbf{x}_2) - \rho_2(\mathbf{x}_3 + \boldsymbol{\Delta}, \mathbf{x}_2)) f_{13}\, dx_3.$$

Since $\boldsymbol{\Delta}$ is arbitrary, the above equation must have a nonzero solution in a stable solid. This line of reasoning can be extended to all the equations in (II.2) and in fact to the entire solution to any order. It is instructive to

carry this out to the next order. For $\rho_3(x_1, x_2, x_3)$ we have

$$\rho_3(x_1, x_2, x_3) = z(1 + f_{12})(1 + f_{13})[\rho_2(x_2, x_3) + \int \rho_3(x_4, x_2, x_3) f_{14} dx_4].$$

We rotate particles 1 and 3 about 2 so that

$$\rho_3(\mathbf{x}_1 + \Delta, \mathbf{x}_2, \mathbf{x}_3 + \Delta) = z(1 + f_{12})(1 + f_{13})[\rho_2(\mathbf{x}_2, \mathbf{x}_3 + \Delta) + \int \rho_3(\mathbf{x}_4 + \Delta, \mathbf{x}_2, \mathbf{x}_3 + \Delta) f_{14} dx_4],$$

where all particle distances are held invariant. Since $\rho_2(x_2, x_3)$ is the solution of the above two particle equation, it is invariant under the rotation. The necessity of a nonzero solution to

$$\rho_3(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3) - \rho_3(\mathbf{x}_1 + \Delta, \mathbf{x}_2, \mathbf{x}_3 + \Delta) = z(1 + f_{12})(1 + f_{13}) \int (\rho_3(\mathbf{x}_4, \mathbf{x}_2, \mathbf{x}_3) - \rho_3(\mathbf{x}_4 + \Delta, \mathbf{x}_2, \mathbf{x}_3 + \Delta)) f_{14} d\mathbf{x}_4$$

follows immediately.

To see that the same thing holds true to second order for $\rho_2(\mathbf{x}_1, \mathbf{x}_2)$, one need only note that

$$\int \rho_3(\mathbf{x}_4, \mathbf{x}_2, \mathbf{x}_3) f_{13} f_{14} d\mathbf{x}_3 d\mathbf{x}_4$$

is invariant under a rotation of X_1 about X_2 as long as $\rho_3(\mathbf{x}_4, \mathbf{x}_2, \mathbf{x}_3)$ is a solution of the above equation for $\rho_3(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3)$ (with the obvious variable change) which has the invariance property.

To recapitulate, we have shown that an additional criterion to the Kirkwood is necessary in order to have a stable solid. This criterion is that at least one of the equations

$$\rho_N(x_1, \ldots, x_N) = \prod_{j=2}^{N} (1 + f_{1j}) \int \rho_N(x_{N+1}, x_2 \cdots x_N) f_{1,N+1} dx_{N+1} \tag{IV.1}$$

must have a nonzero solution. We now show that this is not possible for $z = z_0$.

From Eq. (III.5) we have

$$(1 - z\hat{f}(|\mathbf{K}|))\, \rho_0(\mathbf{K}, \mathbf{x}_2 \cdots \mathbf{x}_N) = -\hat{\rho}_I(\mathbf{K}, \mathbf{x}_2 \cdots \mathbf{x}_N).$$

There are two possibilities. Either $\hat{\rho}_I(\mathbf{K}, \mathbf{x}_2 \cdots \mathbf{x}_N)$ is identically zero or it is not. If it is not, the proof proceeds identically to that in Sec. 3. In this case, however, we have the possibility that $\hat{\rho}_I(\mathbf{K}, \mathbf{x}_2 \cdots \mathbf{x}_N)$ is identically 0 and that $\rho_0(\mathbf{K}, \mathbf{x}_2 \cdots \mathbf{x}_N)$ is nonzero only on the point support $|\mathbf{K}| = K_\alpha$, where

$$1 - z\hat{f}(K_\alpha) = 0.$$

This implies, however, that the Fourier transform of $\hat{\rho}_0(\mathbf{K}, \mathbf{x} \cdots \mathbf{x}_N)$ is an analytic function of $\mathbf{x}_1$. This follows from a theorem in the theory of distributions which states that the Fourier transform of a distribution defined on bounded support is analytic.[4] This is clearly not possible in this case due to the factor

$$\prod_{j=2}^{N} (1 + f_{1j})$$

unless the function is identically zero. This concludes the proof. There exist now two possibilities. Either the values Z_c is a limit point of the eigenvalue spectrum in which case it would indicate an instability of the solid phase, or the spectrum starts at a higher Z in which case the phase transition to a stable solid would be at a higher Z. As this result is identical to Kirkwood's, the former is expected to be correct although no proof is available at present.

V. ONE-DIMENSIONAL CASE

It is interesting to see what this theory predicts in one dimension for hard rods. Since this is a perturbation theory, it is not consistent with the assumption of convergence that the symmetry of each individual term not be the symmetry of the solid. With this in mind we note that for rods of length $2L$

$$\hat{f}(K) = (\sin KL)/K.$$

We require $z > 0$ so that for

$$1 - z\hat{f}(K) = 0,$$

$$\sin KL < 0$$

so that

$$KL > \pi, \quad KL = (2\pi/\lambda)L > \pi, \quad \lambda < 2L.$$

Therefore

$$\rho_1(x_1) = z/[1 - z\hat{f}(0)] + \sum_\alpha A_\alpha [\exp(i\mathbf{K}_\alpha \mathbf{x}_1) \pm \exp(-i\mathbf{K}_\alpha \mathbf{x}_1)].$$

Since $\rho_1(x_1)$ must have the symmetry of the solid, the $\mathbf{K}_\alpha$ are reciprocal lattice vectors and hence λ is the lattice constant in one dimension. But $\lambda < 2L$, which means this solution is not physical. In this regard it is interesting to note that a nonphysical solution for hard rods was found by Gallavotti and Lebowitz.[5,6]

VI. RESULTS AND CONCLUSIONS

We have obtained a result identical in form to Kirkwood's. This result is known to be incorrect as a prediction of the hard sphere freezing transition. However, as we have shown, this condition seems to indicate an instability. It has been shown, however,[7] that a rigorous statistical mechanical treatment of the distribution functions and the free energy show no such instability. However, this approach and Kirkwood's have in common the omission of higher order correlation functions. This is shown clearly in Ref. 7 as the inclusion of terms in eliminate the instability of Kirkwood. It is of interest, however, to consider the above treatment and Kirkwood's as a "mean field" type of approach which might show spinodal points, such as appear in the work of Hoover and Ree,[8] as instabilities. This has been done and will be reported in a future communication.

The value for Z_c in the units chosen is 11.6. As the machine calculations indicate a much higher Z for the phase transition, it seems that the series to be useful must be renormalized.

ACKNOWLEDGMENT

I would like to acknowledge the help of Dr. S. Krebs in preparing the manuscript.

[1]W. Klein, J. Math. Phys. **14**, 1049 (1973).
[2]J. G. Kirkwood, in *Symposium on Phase Transformations in Solids, August* 1948, edited by R. Smoluchowski (Wiley, New York, 1957).
[3]D. Ruelle, *Statistical Mechanics, Rigorous Results* (Benjamin, New York, 1969).
[4]L. Schwartz, *Mathematics for the Physical Sciences* (Addison-Wesley, Reading, Mass., 1966), p. 189.
[5]G. Gallovotti and J. L. Lebowitz, Physica **70**, 219 (1973).
[6]H. J. Brascamp, Commun. Math. Phys. **40**, 235 (1975).
[7]W. Kunkin and H. L. Frisch, J. Chem. Phys. **50**, 1817 (1969).
[8]W. Hoover and F. Ree, J. Chem. Phys. **49**, 3609 (1968).

Rigorous derivation of the Kirkwood–Monroe equation for small activity

N. Grewe and W. Klein

Institut für Theoretische Physik der Universität zu Köln, 5 Köln 41, Zülpicher Str. 77, Germany
(Received 13 June 1975)

We show, for small values of z, that the solution of the Kirkwood–Salsburg equation approaches, in the norm topology, the solution of the Kirkwood–Monroe and van Kampen equations if the potential of interaction is the Kac potential $\phi(x_{12}) = \gamma^s g(\gamma x_{12})$ and the limit $\gamma \to 0$ is taken. We have to assume that the function g is bounded and absolutely integrable and that $\Sigma_{i \neq j} g(\gamma x_{ij}) \geq -mB$ $(B < \infty)$, the sum being performed over all pairs of the m particles.

1. INTRODUCTION

In 1941 Kirkwood and Monroe[1] introduced a theory of freezing. The main result of the theory is that the density distribution obeys an integral equation of the form[2]

$$\rho(\mathbf{x}) = C \cdot \exp[-\beta \int d^s\mathbf{x}' \rho(\mathbf{x}')K(\mathbf{x}-\mathbf{x}')] \tag{1.1}$$

where C is a constant, s the dimensionality of the system and $\beta = 1/k_B T$. Arguments have been given[1,3] that indicate that Eq. (1.1) has periodic solutions which are taken to represent the crystalline phase. The above-cited authors, however, made no attempt to place (1.1) on a rigorous basis.

Gates[4] has presented an argument that Eq. (1.1) can be derived in the limit $\gamma \to 0$ from a variational principle developed by Gates and Penrose[5,6] in the special case where the potential of interaction has the form

$$\phi(\mathbf{x}_2 - \mathbf{x}_1) \equiv \phi(x_{12}) = \gamma^s g(\gamma x_{12}), \tag{1.2}$$

where g consists of two parts g_1 and g_2 with

$$g_1(x) \geq 0, \quad \int_{\mathbb{R}^s} d^s\mathbf{x} \exp(-i\mathbf{k}\mathbf{x}) g_2(x) \geq 0 \quad \text{for all } \mathbf{k}. \tag{1.2'}$$

This is the well-known Kac potential.

However, as Gates pointed out, the above-mentioned variational principle has not been rigorously established for such potentials. Moreover, even if one assumes the existence of such a principle, the derivation of (1.1) is not rigorous.

The results that we wish to communicate were inspired by another result in the above-quoted paper of Gates.[4] He showed that the solution of (1.1) would, for small enough density, be a solution of the following set of linear equations:

$$\begin{aligned}\rho_m(\mathbf{x}_1, \ldots, \mathbf{x}_m) \\ z = [\delta_{m,1} + (1-\delta_{m,1})\rho_{m-1}(\mathbf{x}_2, \ldots, \mathbf{x}_m) \\ + \sum_{n=1}^{\infty} \frac{1}{n!} \int_{\mathbb{R}^{ns}} d^s\mathbf{x}_{m+1} \cdots d^s\mathbf{x}_{m+n} \\ \times \rho_{m+n-1}(\mathbf{x}_2, \ldots, \mathbf{x}_{m+n}) K_n(\mathbf{x}_1; \mathbf{x}_{m+1}, \ldots, \mathbf{x}_{m+n})]\end{aligned} \tag{1.3}$$

with

$$K_n(\mathbf{x}_1; \mathbf{x}_{m+1}, \ldots, \mathbf{x}_{m+n}) = (-\beta)^n \prod_{j=m+1}^{m+n} g(x_{1j})$$

and z being the activity, if one defines the higher order distribution functions by

$$\rho_m(\mathbf{x}_1, \ldots, \mathbf{x}_m) = \prod_{j=1}^{m} \rho(\mathbf{x}_j). \tag{1.4}$$

Equation (1.3) bears a strong resemblance to the Kirkwood–Salsburg[7] equations. They can be written concisely as

$$\boldsymbol{\rho} = z\boldsymbol{\alpha} + z\hat{Q}_0\boldsymbol{\rho}, \tag{1.5}$$

where $\boldsymbol{\rho} \equiv (\rho_1(\mathbf{x}_1), \rho_2(\mathbf{x}_1, \mathbf{x}_2), \cdots)$, $\boldsymbol{\alpha} \equiv (1, 0, 0, \cdots)$ and the effect of $\hat{Q}_0$ on $\boldsymbol{\rho}$ is given by (1.3).

We show, for those values of z and β with $|z| \leq z_1(\beta)$, z_1 depending on bounds for the function g_1, that the solution of the Kirkwood–Salsburg equation in the case of $g_2 \equiv 0$ approaches the solution of (1.3) as $\gamma \to 0$ in the topology generated by the vector norm (2.5). In addition it is shown how the proof can be extended for a more general potential (1.2).

This establishes that the Kirkwood–Monroe equation has, in this range of z and β, a rigorous foundation in statistical mechanics. We also indicate how one can generate, for potentials of this type, a power series in γ^s for the distribution functions.[9]

2. RESULTS CONCERNING THE OPERATORS

We consider a system where particles interact via a Kac potential of the form (1.2). For the function g we demand that

$$\begin{aligned}&0 \leq g(x) \leq A < \infty, \\ &\int_{\mathbb{R}^s} d^s\mathbf{x}\, g(x)^\nu = C_\nu < \infty \quad (\nu = 1, 2).\end{aligned} \tag{2.1}$$

The corresponding Mayer function is

$$f(x) = \exp[-\beta\phi(x)] - 1, \quad \beta = 1/k_B T. \tag{2.2}$$

Property (2.1) implies the regularity of the pair potential ϕ[8]:

$$C(\beta) := \int_{\mathbb{R}^s} d^s\mathbf{x} \,|f(x)| \leq \beta C_1 < \infty. \tag{2.3}$$

This is easily seen from $0 \leq 1 - \exp(-x) \leq x$ for $x \geq 0$.

The Kirkwood–Salsburg operator is defined by

$$\begin{aligned}(\hat{K}\boldsymbol{\varphi})_1(\mathbf{x}_1) = \sum_{n=1}^{\infty} \frac{1}{n!} \int_{\mathbb{R}^{ns}} \varphi_n(\mathbf{x}_2, \ldots, \mathbf{x}_{n+1}) \\ \times \prod_{j=2}^{n+1} f(x_{1j})\, d^s\mathbf{x}_j,\end{aligned}$$

$$(\hat{K}\varphi)_m(\mathbf{x}_1, \ldots, \mathbf{x}_m) = \prod_{j=2}^{m}(1+f(x_{1j}))[\varphi_{m-1}(\mathbf{x}_2, \ldots, \mathbf{x}_m) + \sum_{n=1}^{\infty}\frac{1}{n!} \times \int_{\mathbb{R}^{ns}} \varphi_{m+n-1}(\mathbf{x}_2, \ldots, \mathbf{x}_{m+n}) \prod_{j=m+1}^{m+n} f(x_{1j})\, d^s\mathbf{x}_j] \quad (m \geq 2). \tag{2.4}$$

It acts on the Banach space E_ξ of sequences of functions $\varphi = (\varphi_1, \varphi_2, \cdots)$, where φ_n: $\mathbb{R}^{ns} \to \mathbb{C}$ is Lebesgue-measurable and bounded, the norm being

$$\|\varphi\|_\xi = \sup_{n\in\mathbb{N}} \{\operatorname{ess\,sup}_{(\mathbf{x}_1,\ldots,\mathbf{x}_n)\in\mathbb{R}^{ns}} [|\varphi_n(\mathbf{x}_1, \ldots, \mathbf{x}_n)|/\xi^n]\}, \quad \xi > 0 \text{ fixed.} \tag{2.5}$$

It has been shown by Ruelle[8] that the Kirkwood–Salsburg operator is bounded in the operator norm $\|\ \|$, corresponding to the vector norm (2.5), in the special case of our potential by

$$\|\hat{K}\| \leq (1/\xi)\exp[\xi C(\beta)] =: z_0(\beta)^{-1}. \tag{2.6}$$

We will get this result as a consequence of Lemma 1.

In the next section we will be concerned with solutions of the Kirkwood–Salsburg hierarchy of equations:

$$(\hat{I} - z\hat{K})\kappa = z\alpha, \quad \hat{I} = \text{Unit operator.} \tag{2.7}$$

Our general aim is to show that the Kirkwood–Salsburg equations in certain limiting cases can be approximated by a simpler hierarchy. A special difficulty arises from the product of $[1+f(x_{1j})]$-factors in the definition (2.4) of $\hat{K}$. Depending on the choice of arguments the effect of $\Pi_{j=2}^{m}[1+f(x_{1j})]$ can become very large for high enough m. Therefore, we first introduce a decomposition of $\hat{K}$,

$$\hat{K} = \hat{P}\hat{Q}, \tag{2.8}$$

where the operator $\hat{P}$ provides the multiplication with this product,

$$(\hat{P}\varphi)_1(\mathbf{x}_1) = \varphi_1(\mathbf{x}_1),$$
$$(\hat{P}\varphi)_m(\mathbf{x}_1, \ldots, \mathbf{x}_m) = (\prod_{j=2}^{m}[1+f(x_{1j})])\varphi_m(\mathbf{x}_1, \ldots, \mathbf{x}_m), \tag{2.8a}$$

and $\hat{Q}$ contains the rest of $\hat{K}$:

$$(\hat{Q}\varphi)_m(\mathbf{x}_1, \ldots, \mathbf{x}_m) = (1-\delta_{m,1})\varphi_{m-1}(\mathbf{x}_2, \ldots, \mathbf{x}_m) + \sum_{n=1}^{\infty}\frac{1}{n!}\int_{\mathbb{R}^{ns}} \varphi_{n+m-1}(\mathbf{x}_2, \ldots, \mathbf{x}_{m+n}) \prod_{j=m+1}^{m+n} f(x_{1j})\, d^s\mathbf{x}_j. \tag{2.8b}$$

It is easy to give bounds for these two operators:

Lemma 1: (1) $\|\hat{P}\| \leq 1$, (2) $\|\hat{Q}\| \leq z_0^{-1}$.

Proof: Part (1) follows immediately from the fact that, for all $x \geq 0$, $0 \leq 1+f(x) \leq 1$. With the use of $|\varphi_m(\mathbf{x}_1, \ldots, \mathbf{x}_m)| \leq \|\varphi\|_\xi \cdot \xi^m$ ($\varphi \in E_\xi$) and of (2.3) we have

$$|(\hat{Q}\varphi)_m(\mathbf{x}_1, \ldots, \mathbf{x}_m)| \leq \|\varphi\|_\xi \cdot \xi^{m-1}\left(1-\delta_{m,1}| + \sum_{n=1}^{\infty}\frac{\xi^n}{n!}\left|\int_{\mathbb{R}^s} f(x)\, d^s\mathbf{x}\right|^n\right) \leq \|\varphi\|_\xi \cdot \xi^{m-1}\exp[\xi C(\beta)].$$

This implies $\|\hat{Q}\| \leq (1/\xi)\exp[\xi C(\beta)] = z_0^{-1}$.

N.B.: Since $\|\hat{K}\| \leq \|\hat{P}\| \cdot \|\hat{Q}\| \leq \|\hat{Q}\|$, $\|\hat{K}\| \leq z_0^{-1}$. QED

In practical calculations it will be necessary to truncate the Kirkwood–Salsburg hierarchy, which is of infinite order. This concept is also useful as a tool in our considerations.

The truncation can be described by use of a projection operator

$$\hat{T}_{m_0}\varphi = (\varphi_1, \varphi_2, \ldots, \varphi_{m_0}, 0, 0, \cdots) \tag{2.9}$$

with $m_0 \in \mathbb{N}$ fixed which projects onto the Banach subspace $E_\xi^{(m_0)}$ of all sequences of functions out of E_ξ, whose components with order higher than m_0 are zero. Operators $\hat{O}$ can be confined to this subspace in the following way:

$$\hat{O}^{(m_0)} = \hat{T}_{m_0}\hat{O}\hat{T}_{m_0}. \tag{2.10}$$

A vector $\varphi \in E_\xi$ whose belonging to $E_\xi^{(m_0)}$ shall be stated explicitly will be marked as $\varphi^{(m_0)}$.

As it is possible to expand the Mayer functions in the restricted Kirkwood–Salsburg operator $\hat{Q}$ in powers of $\gamma^s g(\gamma x_{1j})$ we expect a series representation of the form

$$\hat{Q} = \sum_{\nu=0}^{\infty}\gamma^{\nu s}\hat{Q}_\nu, \tag{2.11}$$

where the operators $\hat{Q}_\nu$ are bounded, their norm being independent of γ. This last property arises from the fact that the γ dependence of the $\hat{Q}_\nu$ can be eliminated by rescaling the arguments of vectors $\varphi \in E_\xi$ because of the special form of the Kac potential.

In the following we restrict ourselves to the zeroth order approximation:

$$(\hat{Q}_0\varphi)_m(\mathbf{x}_1, \ldots, \mathbf{x}_m) = (1-\delta_{m,1})\varphi_{m-1}(\mathbf{x}_2, \ldots, \mathbf{x}_m) + \sum_{n=1}^{\infty}\frac{(-\beta\gamma^s)^n}{n!}\int_{\mathbb{R}^{ns}} \varphi_{m+n-1}(\mathbf{x}_2, \ldots, \mathbf{x}_{m+n}) \prod_{j=m+1}^{m+n} g(\gamma x_{1j})d^s\mathbf{x}_j. \tag{2.12}$$

By rescaling the arguments $\mathbf{x}_1, \ldots, \mathbf{x}_m$ and changing the variables of integration with γ, it is obvious that $\hat{Q}_0$ goes over into that one defined in (1.3), (1.5) acting on the rescaled space E_ξ.

The norm of $\hat{Q}_0$ can be bounded as follows:

Lemma 2: $\|\hat{Q}_0\| \leq z_1^{-1}$ with $z_1(\beta) = \xi\exp[-\xi\beta C_1] \leq z_0(\beta)$.

Proof: The proof of $\|\hat{Q}_0\| \leq z_1^{-1}$ is analogous to that of Lemma 1 using (2.1) instead of (2.3). (2.3) then proves $z_1 \leq z_0$. QED

It remains to show that the difference between $\hat{Q}$ and $\hat{Q}_0$ is of order γ^s in the norm:

Theorem 1: $\|Q - \hat{Q}_0\| \leq \gamma^s \cdot \frac{1}{2}\beta^2 C_2 \xi z_1^{-1}$.

Proof: We first note that the difference of the two nonnegative constants βC_1 and $C(\beta)$ can be bounded by

$$0 \leq \beta C_1 - C(\beta) = \int_{\mathbb{R}^s} d^s\mathbf{x}\{\beta\gamma^s g(\gamma x) + \exp[-\beta\gamma^s g(\gamma x)] - 1\}$$
$$\leq \tfrac{1}{2}\gamma^{2s}\beta^2 \int_{\mathbb{R}^s} d^s\mathbf{x}\, g(\gamma x)^2 = \tfrac{1}{2}\gamma^s\beta^2 C_2.$$

Therefore,

$$\left|((\hat{Q} - \hat{Q}_0)\varphi)_m(\mathbf{x}_1, \ldots, \mathbf{x}_m)\right|$$
$$\leq \sum_{n=1}^{\infty} \frac{1}{n!} \left| \int_{\mathbb{R}^{ns}} d^s\mathbf{x}_{m+1} \cdots d^s\mathbf{x}_{m+n} \right.$$
$$\times \Big(\prod_{j=m+1}^{m+n} [\beta\gamma^s g(\gamma x_{1j})] - \prod_{j=m+1}^{m+n} [-f(x_{1j})] \Big)$$
$$\left. \times \varphi_{m+n-1}(\mathbf{x}_2, \ldots, \mathbf{x}_{m+n}) \right|$$
$$\leq \|\varphi\|_\xi \cdot \xi^m \sum_{n=1}^{\infty} \frac{\xi^{n-1}}{n!} [(\beta C_1)^n - C(\beta)^n]$$
$$= \|\varphi\|_\xi \cdot \xi^m [\beta C_1 - C(\beta)] \cdot \sum_{n=1}^{\infty} \frac{\xi^{n-1}}{n!}$$
$$\times \sum_{\nu=0}^{n-1} (\beta C_1)^\nu C(\beta)^{n-1-\nu}$$
$$\leq \|\varphi\|_\xi \cdot \xi^m \gamma^s \cdot \tfrac{1}{2}\beta^2 C_2 \sum_{n=1}^{\infty} \frac{\xi^{n-1}}{(n-1)!} (\beta C_1)^{n-1}$$
$$= \|\varphi\|_\xi \cdot \xi^m \gamma^s \cdot \tfrac{1}{2}\beta^2 C_2 \exp(\xi\beta C_1).$$

This implies $\|\hat{Q} - \hat{Q}_0\| \leq \gamma^s \cdot \frac{1}{2}\beta^2 C_2 \exp(\xi\beta C_1)$. QED

We now want to compare $\hat{Q}_0$ directly to the original Kirkwood–Salsburg operator $\hat{K}$. To eliminate the difficulty with the operator $\hat{P}$, we have to confine ourselves to the subspace $E_\xi^{(m_0)}$, where m_0 is arbitrary but fixed. Then we are able to give a bound to the difference between $\hat{P}^{(m_0)}$ and the unit operator $\hat{I}^{(m_0)}$:

Lemma 3: $\|\hat{P}^{(m_0)} - \hat{I}^{(m_0)}\| \leq \gamma^s\beta A(m_0 - 1)$.

Proof: $((\hat{P}^{(m_0)} - \hat{I}^{(m_0)})\varphi)_m(\mathbf{x}_1, \ldots, \mathbf{x}_m) = 0$ for $m = 1$ and $m > m_0$. For $2 \leq m \leq m_0$ we have with $0 \leq 1 - \exp(-x) \leq x$ $(x \geq 0)$:

$$\left|((\hat{P}^{(m_0)} - \hat{I}^{(m_0)})\varphi)_m(\mathbf{x}_1, \ldots, \mathbf{x}_m)\right|$$
$$\leq \|\varphi\|_\xi \cdot \xi^m \cdot \left| \prod_{j=2}^{m} [1 + f(x_{1j})] - 1 \right|$$
$$= \|\varphi\|_\xi \cdot \xi^m \cdot \left| \exp\Big(-\beta \sum_{j=2}^{m} \phi(x_{1j})\Big) \right.$$
$$\left. - 1 \right| \leq \|\varphi\|_\xi \cdot \xi^m \beta \sum_{j=2}^{m} \left|\phi(x_{1j})\right|$$
$$\leq \|\varphi\|_\xi \cdot \xi^m \gamma^s \beta A(m_0 - 1).$$

Now the above inequality is easily seen. QED

With the help of Theorem 1 and Lemma 3 we can estimate the difference between $\hat{K}^{(m_0)}$ and $\hat{Q}_0^{(m_0)}$:

Theorem 2: $\|\hat{K}^{(m_0)} - \hat{Q}_0^{(m_0)}\| \leq \gamma^s\beta[A(m_0 - 1) + \frac{1}{2}\xi\beta C_2] z_1^{-1}$.

Proof:

$$\|\hat{K}^{(m_0)} - \hat{Q}_0^{(m_0)}\| = \|\hat{P}^{(m_0)}\hat{Q}^{(m_0)} - \hat{Q}_0^{(m_0)}\|$$
$$\leq \|\hat{P}^{(m_0)}\hat{Q}^{(m_0)} - \hat{Q}^{(m_0)}\| + \|\hat{Q}^{(m_0)} - \hat{Q}_0^{(m_0)}\|$$
$$\leq \|\hat{P}^{(m_0)} - \hat{I}^{m_0)}\| \cdot \|\hat{T}_{m_0}\| \cdot \|\hat{Q}\| \cdot \|\hat{T}_{m_0}\|$$
$$+ \|\hat{T}_{m_0}\| \cdot \|\hat{Q} - \hat{Q}_0\| \cdot \|\hat{T}_{m_0}\|.$$

Here we have used $\hat{T}_{m0}^2 = \hat{T}_{m0}$. Clearly the norm of the projector $\hat{T}_{m_0}$ is equal to 1. With this, Lemmas 1, 3, and Theorem 1 we finally have

$$\|\hat{K}^{(m_0)} - \hat{Q}_0^{(m_0)}\| \leq \gamma^s\beta A(m_0 - 1) z_0^{-1} + \gamma^s \cdot \tfrac{1}{2}\beta^2 C_2 \xi z_1^{-1}.$$

The theorem now follows from $z_1 \leq z_0$. QED

3. RESULTS CONCERNING SOLUTIONS OF THE HIERARCHIES

We first ask for the consequences of Theorem 2 concerning solutions $\kappa^{(m_0)}$ and $\rho_0^{(m_0)}$ of the corresponding truncated hierarchies:

$$(\hat{I}^{(m_0)} - z\hat{K}^{(m_0)})\kappa^{(m_0)} = z\alpha,$$
$$(\hat{I}^{(m_0)} - z\hat{Q}_0^{(m_0)})\rho_0^{(m_0)} = z\alpha. \tag{3.1}$$

From these two equations we have the following property of $\kappa^{(m_0)}$ and $\rho_0^{(m_0)}$:

Lemma 4: $(\hat{I}^{(m_0)} - z\hat{K}^{(m_0)})\,(\kappa^{(m_0)} - \rho_0^{(m_0)}) = z(\hat{K}^{(m_0)} - \hat{Q}_0^{(m_0)})\rho_0^{(m_0)}$.

Proof: The lemma is easily proven by subtracting the second equation from the first one and adding a term $z\hat{K}^{(m_0)}\rho_0^{(m_0)}$ on each side. QED

Theorem 2 and Lemma 4 enable us to make a statement about the quality of $\rho_0^{(m_0)}$ as a solution of the truncated Kirkwood–Salsburg equations in dependence of the parameter γ:

Theorem 3: $\|(\hat{I}^{(m_0)} - z\hat{K}^{(m_0)})(\kappa^{(m_0)} - \rho_0^{(m_0)})\|_\xi \leq \gamma^s |z| M_{m_0}$. with $M_{m_0} = \beta[A(m_0 - 1) + \frac{1}{2}\xi\beta C_2] \cdot z_1^{-1} \cdot \|\rho_0^{(m_0)}\|_\xi$ independent of γ.

Proof: Lemma 4 and Theorem 2 give

$$\|(\hat{I}^{(m_0)} - z\hat{K}^{(m_0)})(\kappa^{(m_0)} - \rho_0^{(m_0)})\|_\xi$$
$$\leq |z| \cdot \|\hat{K}^{(m_0)} - \hat{Q}_0^{(m_0)}\| \cdot \|\rho_0^{(m_0)}\|_\xi \leq \gamma^s |z| M_{m_0}.$$

It remains to be shown that $\|\rho_0^{(m_0)}\|_\xi$ does not depend on γ. But this results from the fact that the dependence of $\hat{Q}_0$ (and therefore of $\hat{Q}_0^{(m_0)}$) on γ can be eliminated by rescaling $\rho_0^{(m_0)}$, which does not change its norm. QED

Theorem 3 states that for small γ the difference $\kappa^{(m_0)} - \rho_0^{(m_0)}$ is close to the kernel of the operator $\hat{I}^{m_0)} - z\hat{K}^{(m_0)}$, so that $\rho_0^{(m_0)}$ is in the neighborhood of a solution of the truncated Kirkwood–Salsburg equations.

Additional information can be derived inside the circle $|z| < z_1$. The reason for this is that the operators $\hat{K}$, $\hat{K}^{(m_0)}$, $\hat{Q}_0$, $\hat{Q}_0^{(m_0)}$ all have norm less or equal z_1^{-1}. Therefore, the following Lemma can be used:

Lemma 5: Let $\hat{O}$ be a linear operator on E_ξ with $\|\hat{O}\| \leq z_1^{-1}$. Let ω and $\omega^{(m_0)}$ be solutions of the equations $(\hat{I} - z\hat{O})\,\omega = z\alpha$ and $(\hat{I}^{(m_0)} - z\hat{O}^{(m_0)})\omega^{(m_0)} = z\alpha$, respectively,

with $|z| < z_1$. Then:

$$\|\omega - \omega^{(m_0)}\|_\xi \leq 2\frac{z_1}{\xi}\frac{(|z|/z_1)^{m_0+1}}{1-|z|/z_1}.$$

Proof: Using Neumann's theorem, we know that $(\hat{I} - z\hat{O})$ and $(\hat{I}^{m_0} - z\hat{O}^{(m_0)})$ have both an inverse on E_ξ and $E_\xi^{(m_0)}$ respectively, which are given by

$$(\hat{I} - z\hat{O})^{-1} = \sum_{\nu=0}^{\infty} (z\hat{O})^\nu,$$
$$(\hat{I}^{(m_0)} - zO^{(m_0)})^{-1} = \sum_{\nu=0}^{\infty} (z\hat{O}^{(m_0)})^\nu.$$

Since $z\boldsymbol{\alpha} \in E_\xi^{(m_0)}$, ω, and $\omega^{(m_0)}$ are both uniquely determined, their difference can be bounded in the following way:

$$\|\omega - \omega^{(m_0)}\|_\xi \leq |z| \cdot \Big(\sum_{\nu=0}^{m_0-1} \|(z\hat{O})^\nu \boldsymbol{\alpha} - (z\hat{O}^{(m_0)})^\nu \boldsymbol{\alpha}\|_\xi + \sum_{\nu=m_0}^{\infty} \|(z\hat{O})^\nu \boldsymbol{\alpha} - (z\hat{O}^{(m_0)})^\nu \boldsymbol{\alpha}\|_\xi\Big).$$

In the first sum each term gives zero. This is because only the first component of $\boldsymbol{\alpha}$ is different from zero and in $\hat{O}^\nu \boldsymbol{\alpha}$ only the first $\nu+1$ components, so that all vectors in the first sum are in $E_\xi^{(m_0)}$. Since $\hat{O}$ and $\hat{O}^{(m_0)}$ are identical on $E_\xi^{(m_0)}$, we have the desired result. Then

$$\begin{aligned}
&\|\omega - \omega^{(m_0)}\|_\xi \\
&\leq |z| \cdot \sum_{\nu=m_0}^{\infty} (\|z\hat{O}\|^\nu + \|z\hat{O}^{(m_0)}\|^\nu)\|\boldsymbol{\alpha}\|_\xi \\
&\leq 2|z| \sum_{\nu=m_0}^{\infty} \left(\frac{|z|}{z_1}\right)^\nu \cdot \frac{1}{\xi} \\
&= 2\frac{|z|}{\xi}\left(\frac{|z|}{z_1}\right)^{m_0} \frac{1}{1-|z|/z_1} \\
&= 2\frac{z_1}{\xi}\frac{(|z|/z_1)^{m_0+1}}{1-|z|/z_1}. \qquad \text{QED}
\end{aligned}$$

Now we are able to give a bound to the difference of the (uniquely determined) solutions of the full Kirkwood–Salsburg equations and the $\hat{Q}_0$ hierarchy inside the circle $|z| < z_1$:

Theorem 4: Let κ and ρ_0 be the uniquely determined solutions of $(\hat{I} - z\hat{K})\kappa = z\boldsymbol{\alpha}$ and $(\hat{I} - z\hat{Q}_0)\rho_0 = z\boldsymbol{\alpha}$, respectively, with $|z| < z_1$. Then for any $m_0 \in \mathbb{N}$,

$$\|\kappa - \rho_0\|_\xi \leq \frac{1}{1-|z|/z_1}\left[4\frac{z_1}{\xi}\left(\frac{|z|}{z_1}\right)^{m_0+1} + \gamma^s |z| M_{m_0}\right]$$

with the same M_{m_0} as in Theorem 3.

Proof: With Theorem 2, Lemmas 4,5, Neumann's theorem, and $\kappa^{(m_0)}, \rho_0^{(m_0)}$ as in (3.1), we have

$$\begin{aligned}
\|\kappa - \rho_0\|_\xi &\leq \|\kappa - \kappa^{(m_0)}\|_\xi + \|\kappa^{(m_0)} - \rho_0^{(m_0)}\|_\xi + \|\rho_0^{(m_0)} - \rho_0\|_\xi \\
&\leq 4\frac{z_1}{\xi}\frac{(|z|/z_1)^{m_0+1}}{1-|z|/z_1} + \|(\hat{I}^{(m_0)} - z\hat{K}^{(m_0)})^{-1}\| \\
&\quad \cdot |z| \cdot \|\hat{K}^{(m_0)} - \hat{Q}_0^{(m_0)}\| \cdot \|\rho_0^{(m_0)}\|_\xi \\
&\leq 4\frac{z_1}{\xi}\frac{(|z|/z_1)^{m_0+1}}{1-|z|/z_1} + \frac{1}{1-|z|/z_1}|z|\gamma^s M_{m_0}. \qquad \text{QED}
\end{aligned}$$

In fact we can make the difference between the two solutions as small as we want by chosing first a large $m_0 \in \mathbb{N}$, which causes the first term to be small, and then using a properly small γ in the second term.

We will conclude this section with some remarks about the hierarchy corresponding to the operator $\hat{Q}_0$. As Gates has pointed out,[4] the equation

$$(\hat{I} - z\hat{Q}_0)\rho_0 = z\boldsymbol{\alpha} \tag{3.2}$$

is reduced by an ansatz of the form

$$\rho_0 = (\rho_1, \rho_2, \ldots),$$
$$\rho_m(\mathbf{x}_1, \ldots, \mathbf{x}_m) = \prod_{j=1}^{m} \rho_1(\mathbf{x}_j) \quad (m \in \mathbb{N}) \tag{3.3}$$

to the Kirkwood–Monroe and van Kampen equation:

$$\rho_1(\mathbf{x}_1) = z\exp[-\beta\int_{\mathbb{R}^s} d^s\mathbf{x}_2 \phi(x_{12})\rho_1(\mathbf{x}_2)]. \tag{3.4}$$

For all positive z we have a constant solution, uniquely determined by

$$\rho_1 = z\exp(-\beta C_1 \rho_1). \tag{3.5}$$

Whether or not there are additional solutions for large enough z containing a sinusoidal term is a question of high physical interest, which is discussed in the paper of Gates,[4] but has remained still unsolved.

Nevertheless, in the circle $|z| < z_1$ the unique solution of the $\hat{Q}_0$ hierarchy is $\rho_0 = (\rho_1, \rho_1^2, \rho_1^3, \cdots)$, where ρ_1 is a solution of (3.5). In fact, this equation for ρ_1 can have more than one solution, for example, in the interval $-(e\beta C_1)^{-1} < z < 0$ there are two. But all of them, except the one with smallest $|\rho_1|$ do not correspond to solutions of the $\hat{Q}_0$ hierarchy. They are ruled out because of $\rho_1/\xi > 1$ which makes the norm of the corresponding vector ρ_0 infinite.

4. EXTENSION OF THE RESULTS AND CONCLUDING REMARKS

In the proofs of the previous sections 2 and 3 we restricted ourselves to positive potentials, i.e., $g_2 \equiv 0$. This was needed to obtain the result that $\Pi_{j=2}^m [1 + f(x_{1j})]$ was bounded. Potentials of the type $\gamma^s g_2(\gamma x_{12})$ do, however, present a more serious problem as they allow

$$\sum_{i \neq j} g_2(\gamma x_{ij}) < -mD, \tag{4.1}$$

where D is an arbitrarily large positive constant and m is the number of particles. These potentials are not stable[8] and can lead to nonthermodynamic behavior. If, however, we restrict ourselves to potentials such that

$$\sum_{i \neq j} g_2(\gamma x_{ij}) \geq -mB,$$
$$\int_{\mathbb{R}^s} d^s\mathbf{x} |g_1(x) + g_2(x)|^\nu = C_\nu' < \infty \quad (\nu = 1, 2), \tag{4.2}$$

where B is a finite positive constant, then the previous proofs can be amended to include such potentials.

This is done by defining, following Ruelle,[8] a permutation operator $\hat{\Pi}$, which permutes the particles such that

$$(\hat{\Pi}\varphi)_m(\mathbf{x}_1, \ldots, \mathbf{x}_m) = \varphi_m(\mathbf{x}_{i_1}, \mathbf{x}_{i_2}, \ldots, \mathbf{x}_{i_m})$$

with

$$\prod_{j=2}^{m}[1+f(x_{i_1 i_j})]<\exp(\gamma^s\beta B)\quad (m\in\mathbb{N}) \tag{4.3}$$

We then have the following:

Theorem 5: In the norm topology the solutions of

$$(\hat{I}-\hat{\Pi}z\hat{K})\kappa=z\alpha \text{ and } (\hat{I}-\hat{\Pi}z\hat{Q}_0)\rho_0=z\alpha \tag{4.4}$$

become arbitrarily close as $\gamma\to 0$ in a circle $|z|<z_2$.

N.B.: (1) One can show, following Ruelle[8] and methods employed in the proof of Lemma 1, that $\|\hat{K}\|$, which, of course, depends on γ, is uniformly bounded in γ in an interval $0<\gamma<\gamma_0$. Then choose: $z_2^{-1}=\sup_{0<\gamma<\gamma_0}\|\hat{K}\|$.

(2) (4.4) clearly implies that ρ_0 corresponds to a solution of the Kirkwood—Monroe and van Kampen equations.

The obvious drawback to the theorems proven above is is that they do not say enough about the solutions of the mean field theory near the freezing transition. The difficulty in proving a theorem similar to Theorem 5 for $|z|\geq z_2$, is that the limit $\gamma\to 0$ of $\prod_{j=2}^{\infty}(1+f(x_{i_1 i_j}))$ is not equal to 1, but is a function of the positions of particles i_j relative to i_1. The proof of such a theorem requires a theory which can, as Ruelle did for small z, state explicitly how the distribution functions of lower order are influenced by the distribution functions $\rho_m(\mathbf{x}_1,\ldots,\mathbf{x}_m)$ in the limit $m\to\infty$.[10]

As mentioned in the Introduction, one can generate from this formalism a power series in γ^s for the distribution functions. One has merely in each term of the Neumann series to expand the Mayer function $f(x_{1j})=\exp[-\beta\gamma^s g(\gamma x_{1j})]-1$ in a power series in $\gamma^s g(\gamma x_{1j})$ and redefine the independent variables as γ^s and γx_{1j} in place of γ and x_{1j}. One actually obtains a double series in z and γ^s. It is clear from Ruelle's work that the error made in truncating the series at a given order of z and γ can be bounded and made arbitrarily small.[11]

[1] J. G. Kirkwood and E. Monroe, J. Chem. Phys. **9**, 514 (1941).

[2] For these potentials the Kirkwood-Monroe and van Kampen equations are, as shown by Gates in Ref. 4, identical. See N. G. van Kampen, Phys. Rev. **135**, A366 (1964).

[3] R. Brout, Physica **29**, 1041 (1963).

[4] D. J. Gates, Ann. Phys. (N.Y.) **71**, 395 (1972).

[5] D. J. Gates and O. Penrose, Commun. Math. Phys. **15**, 253 (1969).

[6] D. J. Gates and O. Penrose, Commun. Math. Phys. **17**, 194 (1970).

[7] J. G. Kirkwood and Z. W. Salsburg, Discuss. Faraday Soc. **15**, 23 (1953).

[8] D. Ruelle, *Statistical Mechanics Rigorous Results* (Benjamin, New York, 1969), Chap. 4.

[9] This was accomplished for potentials of the type $V(x)+\gamma^3 g(\gamma x)$, where $V(x)$ is the hard core potential, by J. L. Lebowitz, G. Stell, and S. Baer, J. Math. Phys. **6**, 1282 (1965).

[10] Nevertheless, it would be desirable to work along these lines, for one knows that in certain cases mean field theory may be exact in the limit $\gamma\to 0$. See F. S. Høye, Phys. Rev. B **9**, 2390 (1974).

[11] Recently, a somewhat different approach to the theory of freezing has been made by Raveché and Stuart.[12] It strongly indicates the possibility of a density distribution function with a periodic structure occuring at the limit of the metastable liquid phase. Work connected with this question is also in progress in the frame of the γ expansions presented in this paper.

[12] H. F. Raveché and C. A. Stuart, J. Chem. Phys. **63**, 1099 (1975).

Spontaneously broken symmetry and cosmological constant

M. Y. Wang*

Center for Theoretical Studies, University of Miami, Coral Gables, Florida
(Received 12 May 1975; revised manuscript received 30 July 1975)

A solution of the Einstein equation with cosmological term produced by spontaneously broken symmetry is presented. The solution implies that the universe will recontract.

In those years the theories of spontaneously broken symmetry and Higgs phenomena have been a topic of active investigation in elementary particle physics.[1] The crucial points are that spontaneously broken symmetry requires a nonzero vacuum expectation value of scalar meson, and the vector meson acquires mass from Higgs mechanism. These mechanisms have been applied to unify the theories of weak, electromagnetic and strong interactions.[1] Recently the question of possible relationships between the spontaneously broken symmetry and the cosmological constant was raised by several authors.[2,3] Their arguments, based on the conjecture of Zeldovich and Novikov[4] are that the vacuum value of energy momentum tensor $T_{\mu\nu}$ appears in the form of a cosmological term in the vacuum field equations.

In this paper, a solution of the Einstein equation with cosmological term produced by spontaneously broken symmetry is presented. The solution is shown to be consistent with the conjecture of Zeldovich and Novikov. The implications of the result are discussed.

Let us consider a system of the triplet scalar and SO(3) gauge fields coupled with the gravitational field. The action of the system can be written as[5]

$$I = I_1 + I_2 + I_3, \tag{1}$$

$$I_1 = -\int \sqrt{-g}\,(R - 2\Lambda)\,d4x, \tag{2}$$

$$I_2 = -\tfrac{1}{4}\int \sqrt{-g}\,G_a^{\mu\nu} G_{\mu\nu a}, \tag{3}$$

$$I_3 = -\tfrac{1}{2}\int \sqrt{-g}\,(D^\mu Q_a D_\mu Q_a) - \tfrac{1}{2}\mu^2 Q_a^2 - \tfrac{1}{8}\lambda (Q_a^2)^2. \tag{4}$$

where Λ is the cosmological constant and

$$G_{\mu\nu} = \partial_\mu W_{\nu a} - \partial_\nu W_{\mu} + e\epsilon_{abc} W_{\mu b} W_{\nu c}, \tag{5}$$

$$D_\mu Q_a = \partial_\mu Q_a + e\epsilon_{abc} W_{\mu b} Q_c, \tag{6}$$

$$G_a^{\mu\nu} = g^{\mu\alpha} g^{\nu\beta} G_{\alpha\beta a}, \tag{7}$$

$$g = \det |g_{\mu\nu}|. \tag{8}$$

W_a and Q_a are a triplet of vector fields and scalar fields respectively. We choose the parameter μ^2 to be negative so that field Q yields a nonzero vacuum expectation value:

$$\langle Q_a \rangle^2 = F^2, \quad \mu^2 = -\lambda F^2/2. \tag{9}$$

Now we ask for a solution of the field equations that is static and spherically symmetric, i.e.,

$$W_{0a} = 0, \quad W_{ia,0} = 0, \quad Q_{a,0} = 0. \tag{10}$$

Following the ansatz of Wu and Yang,[6] we can write Q_a and W_{ia} as

$$Q_a(\mathbf{x}, t) = x^a Q(r), \qquad W_{ia}(\mathbf{x}, t) = \epsilon_{iab} x^b W(r), \tag{11}$$

where ϵ_{iab} is the usual ϵ symbol. The most general static and spherically symmetric tensor $g_{\mu\nu}$ in the cartesian coordinate is shown to be of the form[7]

$$g_{00} = -\alpha(r), \quad g_{0i} = 0, \qquad g_{ij} = \delta_{ij} - (1 - \beta) x^i x^j / r^2, \tag{12}$$

where α and β are function of r only. After some algebra, the Lagrangian L becomes

$$\begin{aligned} L = -4\pi \int r^2 \Big\{ & \alpha^{-1/2}\Big(\frac{d^2\alpha}{dr^2}\Big)\beta^{1/2} + \frac{2}{r}\alpha^{-1/2}\Big(\frac{d\alpha}{dr}\Big)\beta^{-1/2} \\ & - \frac{2}{r}\alpha^{1/2}\Big(\frac{d\beta}{dr}\Big)\beta^{-3/2} + \frac{2}{r^2}\alpha^{1/2}\beta^{-1/2} - \tfrac{1}{2}\alpha^{-3/2}\Big(\frac{d\alpha}{dr}\Big)^2\beta^{-1/2} \\ & - \frac{2}{r^2}\beta^{1/2}\alpha^{1/2} - \tfrac{1}{2}\alpha^{-1/2}\Big(\frac{d\alpha}{dr}\Big)\Big(\frac{d\beta}{dr}\Big)\beta^{-3/2} + \alpha^{1/2}\beta^{1/2} \\ & \times \Big[r^2\Big(\frac{dW}{dr}\Big)^2 + 4rW\Big(\frac{dW}{dr}\Big) + 6W^2 + 2er^2W^3 + \tfrac{1}{2}e^2r^4W^4 \\ & + 4\Big(\frac{1}{\beta} - 1\Big) rW\Big(\frac{dW}{dr}\Big) + \Big(\frac{1}{\beta} - 1\Big) r^2 \\ & \times \Big(\frac{dW}{dr}\Big)^2 + 4\Big(\frac{1}{\beta} - 1\Big) W^2 + \tfrac{3}{2}Q^2 \\ & + rQ\Big(\frac{dQ}{dr}\Big) + 2er^2WQ^2 + \frac{1}{2}\Big(\frac{1}{\beta} - 1\Big) Q^2 + r\Big(\frac{1}{\beta} - 1\Big) Q\Big(\frac{dQ}{dr}\Big) \\ & + \frac{r^2}{2}\Big(\frac{dQ}{dr}\Big)^2 + \frac{1}{2}\Big(\frac{1}{\beta} - 1\Big) r^2\Big(\frac{dQ}{dr}\Big)^2 \\ & + e^2r^4W^2Q^2 - \frac{\lambda F^2}{4} r^2Q^2 + \frac{\lambda}{8} r^4Q^4 - 2\Lambda \Big] \Big\}. \end{aligned} \tag{13}$$

The field equations can be obtained from Eq. (13) by varying α, β, Q, and W. The final forms are

$$\begin{aligned} & r\Big(\frac{d\beta}{dr}\Big) - \beta + \beta^2 \\ & = \frac{r^2}{2}\Big[r^2\Big(\frac{dW}{dr}\Big)^2 + 4rW\Big(\frac{dW}{dr}\Big) + 6W^2 + 2er^2W^3 + \tfrac{1}{2}e^2r^4W^4 \\ & + 4\Big(\frac{1}{\beta} - 1\Big) rW\Big(\frac{dW}{dr}\Big) + \Big(\frac{1}{\beta} - 1\Big) r^2\Big(\frac{dW}{dr}\Big)^2 + 4\Big(\frac{1}{\beta} - 1\Big) W^2 + \tfrac{3}{2}Q^2 \end{aligned}$$

$$+rQ\left(\frac{dQ}{dr}\right)+2er^2WQ^2+\frac{r^2}{2}\left(\frac{dQ}{dr}\right)^2+e^2r^4W^2Q^2+\frac{1}{2}\left(\frac{1}{\beta}-1\right)Q^2$$

$$+\left(\frac{1}{\beta}-1\right)rQ\left(\frac{dQ}{dr}\right)+\frac{1}{2}\left(\frac{1}{\beta}-1\right)r^2\left(\frac{dQ}{dr}\right)^2$$

$$-\frac{\lambda F^2}{4}r^2Q^2+\frac{\lambda}{8}r^4Q^4-2\Lambda\Big]\beta^2, \tag{14}$$

$$r\left(\frac{d\alpha}{dr}\right)+\alpha-\beta\alpha$$

$$=-\frac{r^2}{2}\Big[r^2\left(\frac{dW}{dr}\right)^2+4rW\left(\frac{dW}{dr}\right)+6W^2+2er^2W^3+\tfrac{1}{2}e^2r^4W^4$$

$$+4\left(\frac{1}{\beta}-1\right)rW\left(\frac{dW}{dr}\right)+\left(\frac{1}{\beta}-1\right)r^2\left(\frac{dW}{dr}\right)^2+4\left(\frac{1}{\beta}-1\right)W^2+\tfrac{3}{2}Q^2$$

$$+rQ\left(\frac{dQ}{dr}\right)+2er^2WQ^2+\frac{r^2}{2}\left(\frac{dQ}{dr}\right)^2+e^2r^4W^2Q^2+\frac{1}{2}$$

$$\times\frac{1}{\beta}-1\ Q^2+\left(\frac{1}{\beta}-1\right)rQ\left(\frac{dQ}{dr}\right)+\frac{1}{2}\left(\frac{1}{\beta}-1\right)r^2\left(\frac{dQ}{dr}\right)^2-\frac{\lambda F^2}{4}$$

$$\times r^2Q^2-2\Lambda\ +\frac{\lambda}{8}r^4Q^4\Big]\alpha\beta-r^2\Big[rrW\left(\frac{dW}{dr}\right)+r^2\left(\frac{dW}{dr}\right)^2$$

$$+4W^2+\frac{Q^2}{2}+rQ\left(\frac{dQ}{dr}\right)+\frac{r^2}{2}\left(\frac{dQ}{dr}\right)^2\Big]\alpha, \tag{15}$$

$$\frac{d}{dr}\left\{\left[2\beta^{-1}r^4\left(\frac{dW}{dr}\right)+4\beta^{-1}r^3W\right]\alpha^{1/2}\beta^{1/2}\right\}$$

$$=\Big[4r^2W+6er^4W^2+2e^2r^6W^3+4\beta^{-1}r^3\left(\frac{dW}{dr}\right)$$

$$+8\beta^{-1}r^2W+2er^4Q^2+2e^2r^6WQ^2\Big]\alpha^{1/2}\beta^{1/2}, \tag{16}$$

$$\frac{d}{dr}\left\{\left[\beta^{-1}r^3Q+\beta^{-1}r^4\left(\frac{dQ}{dr}\right)\right]\alpha^{1/2}\beta^{1/2}\right\}$$

$$=\Big[2Qr^2+4er^4WQ+\beta^{-1}Qr^2+r^3\beta^{-1}\left(\frac{dQ}{dr}\right)$$

$$+2e^2r^6W^2Q-\frac{\lambda F^2}{2}r^4Q+\frac{\lambda}{2}r^6Q^3\Big]\alpha^{1/2}\beta^{1/2}. \tag{17}$$

It is easily verified that

$$W(r)=-1/er^2,$$
$$Q(r)=F/r, \tag{18}$$
$$\alpha(r)=\beta^{-1}(r)=1-2m/r+1/4e^2r^2$$

is a solution of Eqs. (9) and (14)–(17). The cosmological constant is found self-consistently to be

$$\Lambda=-\tfrac{1}{16}\lambda F^4 \tag{19}$$

The following remarks on the above solution are in order:

1. Solution (18) reduces to the t'Hooft's magnetic monopole[8] in flat space–time.

2. The cosmological constant Λ is consistent with that of Dreitlein.[3] Thus, if the universe at the present epoch is isotropic, the result indicates that the universe will eventually contract, as has been pointed out by Dreitlein.[3]

Recently Coleman and Weinberg[9] have investigated the possibility that radiative correction may produce spontaneous symmetry broken down. In that case, the radiative correction can be viewed as the dynamical origin of cosmological term. This problem is under investigation.

ACKNOWLEDGMENTS

I am indebted to Dr. Arnold Perlmutter for his cal reading of the manuscript. I would also like to thank Professor Behram Kursunoglu for his hospitality at the Center for Theoretical Studies, Universities of Miami, Carol Gables, Florida.

*Present address, Nuclear Engineering Lab., University of Illinois at Urbana-Champaign, Urbana, Illinois 61801.

[1]G. S. Aber and B. W. Lee, Phys. Rep. C 9, 1 (1973) and references cited therein.
[2]A. D. Linde, Pis. Zh. Eksp. Teor. Fiz. 19, 320 (1974) [JETP Lett. 19, 183 (1974)].
[3]J. Dreitlein, Phys. Rev. Lett. 33, 1243 (1974).
[4]Ya. B. Zeldovich and I. D. Novikov, *Relativistic Astrophysics* (University of Chicago Press, Chicago, Illinois, 1971), Vol. 1, pp. 28ff.
[5]In the context, we have taken the unit $16\pi G=1$.
[6]T. T. Wu and C. N. Yang, *Properties of Matter under Unusual Condition*, edited by Mark and Fernback (Interscience, New York, 1969).
[7]J. L. Anderson, *Principle of Relativity Physics* (Academic, New York, 1967).
[8]G. t'Hooft, Nucl. Phys. B 79, 276 (1974).
[9]S. Coleman and E. Weinberg, Phys. Rev. B 7, 1888 (1973).

Bethe–Salpeter spinor equation at $P_\mu = 0$ and SO(5) spinor spherical harmonics*

E. G. Floratos

Laboratory of Theoretical Physics, Department of Physics, Athens University, Panepistimiopolis, Athens, T.T. 621, Greece
(Received 3 July 1975)

We study the B–S equation, in the ladder approximation, for the zero energy bound states of a spinor and a scalar particle interacting via the exchange of a massless scalar particle. Constructing and using a complete set of SO(5) spinor spherical harmonics, we find the SO(5) degenerate spectrum of the coupling constant and the bound state amplitudes up to a normalization constant. It turns out that the SO(5) symmetry is broken by these amplitudes in a peculiar way.

I. INTRODUCTION

During its long story, the Bethe–Salpeter[1] relativistic covariant equation has been proved[2] a useful theoretical laboratory for attacking, in the framework of the quantum field theory, the two particle interaction problem. Many authors, studying the asymptotic behavior of the em form factors of the nucleon, have considered the nucleon as a bound state of a spinor-scalar particle system, interacting by the exchange of a massless scalar particle and have found good agreement with the experiment.[3–6] However, the above bound state problem has not yet been solved exactly but has been considered only for the asymptotic behavior of the bound state amplitudes. Until now we have analytic solution only for the the spectrum of the coupling constant, for zero energy of the bound states.[7]

In this paper we study in the latter approximation the zero energy bound states of the spinor–scalar particle system. In Sec. II we write down the B–S equation for spinor bound states in momentum space and we transform it using the stereographic projection method.[8] In Sec. III we construct the appropriate for the problem, five-dimensional spinor spherical harmonics which are bases for the $(e=N+\frac{3}{2}, f=\frac{3}{2})$[9] irreducible representation spaces of the SO(5) group. In Sec. IV we find the bound state amplitudes up to a normalization constant and the known spectrum of the coupling constant.

II. B-S EQUATION FOR SPINOR BOUND STATES

We suppose that the interaction between the spinor (spin $\frac{1}{2}$) particle ψ and the scalar particle ϕ is due to the exchange of a massless scalar particle σ that comes from the interaction Lagrangian

$$L_I(x) = \lambda \cdot \bar{\psi}(x)\sigma(x)\psi(x) + g \cdot \phi(x)\sigma(x)\phi(x). \tag{2.1}$$

The corresponding bound state B–S equation in momentum space and in the ladder approximation is the following:

$$[m^2 - (\tfrac{1}{2}p_B + p)^2][m + (\tfrac{1}{2}\not{p}_B - \not{p})]X_{(p;p_B)} = \frac{ia}{\pi^2}\int d^4q \frac{X_{(q;p_B)}}{(q-p)^2 + i\epsilon}, \tag{2.2}$$

where p_B is the bound state four-momentum, $a = \lambda g/16\pi^2$, and we have taken the equal mass case $m_\phi = m_\psi = m$.

Going to the rest system of the bound state and choosing the mass units so that $m=1$, we obtain after the Wick[10] rotation

$$(1 - E_B^2/4 - ip_0E_B + p^2)[1 + (E_B/2 - ip^0)\gamma^0 - i\mathbf{p}\cdot\boldsymbol{\gamma}]X(p;E_B) = (a/\pi^2)\int d^4q X(q;E_B)/(p-q)^2, \tag{2.3}$$

where the metric is now Euclidean $p^2 = p_0^2 + p_1^2 + p_2^2 + p_3^2$ and $\{\gamma^\mu, \gamma^\nu\} = 2\delta^{\mu\nu}$, $\mu, \nu = 0,1,2,3$. If we use SO(4) spinor spherical harmonics,[7,11] we cannot, except when $E_B = 0$, separate even the angular variables from the radial one, because the $(E_B/2)\gamma^0$, ip_0E_0 terms do not commute with the space–time generators of the SO(4) group. Although we study the $E_B = 0$ case, instead of using the SO(4) spinors, and obtain a coupled system of two integral equations for the radial dependence, we shall investigate in what way the spinor character of the problem has changed the known SO(5) symmetry of the scalar–scalar case at $E_B = 0$. So we apply the usual stereographic projection of the four-dimensional Euclidean momentum space on the surface of a five-dimensional sphere unit radius:

$$\eta_\mu = 2p_\mu/(p^2+1), \quad p_\mu = \eta_\mu/(1-\eta_5), \quad \mu = 1,2,3,4 = 0, \qquad \eta_5 = (p^2-1)/(p^2+1), \quad p^2 = (1+\eta_5)/(1-\eta_5). \tag{2.4}$$

In these variables the volume element of the momentum space is connected with the surface element of the sphere by the relation

$$d\Omega_5(\eta) = (1-\eta_5)^4 \cdot d^4p. \tag{2.5}$$

Introducing the above transformation into Eq. (2.3), we obtain for $E_B = 0$

$$(\gamma^5 - \not{\eta})Y(\eta) = (a/4\pi^2)\gamma^5 \int d\Omega_5(\xi)Y(\xi)/(1 - \xi\cdot\eta), \tag{2.6}$$

where

$$Y(\eta) = [(i+\gamma^5)/2]X(p;0)/(1-\eta_5)^3, \tag{2.7}$$

and

$$\not{\eta} = \eta_\alpha\gamma^\alpha, \quad \alpha = 1,2,3,4,5, \quad \gamma^5 = \gamma^1\gamma^2\gamma^3\gamma^4, \quad \gamma^4 = \gamma^0. \tag{2.8}$$

From the form of Eq. (2.6) we see that the appropriate SO(5) representations are the spinor ones and especially those which are obtained from the coupling of the relative orbital angular momentum of the interacting particles and the spin of the spinor particle.

III. SO(5) SPINOR SPHERICAL HARMONICS

In this section we construct the appropriate, for the problem, SO(5) spinor spherical harmonics. As is known,[9] the irreducible unitary representations of the SO(5) group are classified according to the values of two Casimir operators

$$C_1=\tfrac{1}{2}M^{ab}M^{ab} \quad \text{and} \quad C_2=w^a w^a,$$

where

$$w^a=\tfrac{1}{8}\epsilon^{abcde}M^{bc}M^{de}, \tag{3.1}$$

with ϵ^{abcde} the complete antisymmetric tensor of five dimensions, and the M^{ab} are the SO(5) generators which obey the following Lie algebra:

$$[M^{ab},M^{cd}]=i(\delta^{ad}M^{bc}+\delta^{bc}M^{ad}-\delta^{ab}M^{cd}-\delta^{cd}M^{ab}),$$
$$a,b,c,d=1,2,\ldots,5 \tag{3.2}$$

where δ^{ab} is Kronecker delta. The reduction of the SO(5) irreducible representation in SO(4) ones is controlled by two integers or semi-integers simultaneously, e,f such that $e\geq f\geq 1$. This reduction is

$$H_{e,f}=\bigoplus_{(j_1,j_2)\in\Delta} H_{(j_1,j_2)}, \tag{3.3}$$

where Δ is, the following set:

$$\Delta=\{(j_1,j_2)\,|\,:j_1+j_2+1=f,f+1,\ldots,e-1,e,$$
$$j_1-j_2=1-f,2-f,\ldots,f-1\}. \tag{3.4}$$

On these spaces the two Casimir operators C_1,C_2 act as follows:

$$C_1H_{e,f}=[e(e+1)+f(f-1)-2]H_{e,f}, \tag{3.5}$$
$$C_2H_{e,f}=[e(e+1)f(f-1)]H_{e,f}. \tag{3.6}$$

Knowing that the orbital angular momentum in SO(4) language is characterized by the pairs (j,j), $j=n/2$, $n=0,1,2,\cdots$, and the spinor particle in its rest frame by $(\frac{1}{2},0)\oplus(0,\frac{1}{2})$,[11] we see that the representations which concern us are determined by the relation

$$|j_1-j_2|=\tfrac{1}{2}. \tag{3.7}$$

Indeed this relation restricts completely the f to the value $f=\frac{3}{2}$, but the e can take all the values:

$$e=N+\tfrac{3}{2}, \quad N=n_{\max}=0,1,2,\cdots.$$

We construct the bases of the SO(5) irreducible spaces $(f=\frac{3}{2},\ e=N+\frac{3}{2})$, considering the following representation of the M^{ab}:

$$M^{ab}=M^{ab}_{0\tau}+S^{ab},$$

where

$$M^{ab}_{0\tau}=i\left(\eta^a\frac{\partial}{\partial\eta^b}-\eta^b\frac{\partial}{\partial\eta^a}\right), \quad S^{ab}=\frac{i}{4}(\gamma^a\gamma^b-\gamma^b\gamma^a), \tag{3.8}$$
$$a,b=1,2,3,4,5.$$

As bases of the SO(4) irreducible spaces we take those restricted by $|j_1-j_2|=\frac{1}{2}$, that is characterized by the pairs[11] $((n\pm1)/2,n/2)$ and $(n/2,(n\pm1)/2)$. These are known inatwo component formalism.[7,11] Following Rothe,[11] we denote the bases by $Z_{n\pm}$ and $\tilde{Z}_{n\pm}$ respectively, omitting the SO(3) indices. These are functions of the polar angles of the four-dimensional Euclidean space θ_2,θ_1,ϕ:

$$\begin{aligned}&p_4=p\cos\theta_2, \quad p_\mu p_\mu\equiv p^2, \quad \mu=1,2,3,4\\ &p_3=p\sin\theta_2\cos\theta_1,\\ &p_2=p\sin\theta_2\sin\theta_1\sin\phi,\\ &p_1=p\sin\theta_2\sin\theta_1\cos\phi,\end{aligned} \tag{3.9}$$

and they satisfy the following identities[12]:

$$(p_\mu\sigma_\mu)Z_{n\pm}=ip\tilde{Z}_{(n\pm1)\mp}, \tag{3.10}$$
$$(p_\mu\bar{\sigma}_\mu)\tilde{Z}_{n\pm}=-ip\cdot Z_{(n\pm1)\mp}, \tag{3.11}$$
$$(\sigma_\mu\cdot\partial_\mu)Z_{n\pm}=i\tilde{Z}_{(n\pm1)\mp}\left[\frac{d}{dp}\mp\frac{(n+1\mp1)}{p}\right], \tag{3.12}$$
$$(\bar{\sigma}_\mu\cdot\partial_\mu)\tilde{Z}_{n\pm}=-iZ_{(n\pm1)\mp}\left[\frac{d}{dp}\mp\frac{(n+1)\mp1}{p}\right], \tag{3.13}$$

where $\sigma_\mu=(\boldsymbol{\sigma},\sigma_4)$, $\bar{\sigma}_\mu=(\boldsymbol{\sigma},-\sigma_4)$, $\boldsymbol{\sigma}=(\sigma_1,\sigma_2,\sigma_3)$ are the Pauli matrices, $\sigma_4=i$, and $\partial_\mu=\partial/\partial p_\mu$. Now because we are using a four-component formalism we construct the four-component analog of the Z's. These are normalized on the surface of four-dimensional sphere:

$$|n,\pm,j,\mu\rangle_p=\frac{1}{\sqrt{2}}\begin{bmatrix}Z_{n\pm}\\ -Z_{n\pm}\end{bmatrix}, \tag{3.14}$$
$$|n,\pm,j,\mu\rangle^*_p=\frac{1}{\sqrt{2}}\begin{bmatrix}\tilde{Z}_{n\pm}\\ \tilde{Z}_{n\pm}\end{bmatrix}, \tag{3.15}$$

where j,μ are SO(3) indices and p denotes the set of angular variables θ_2, θ_1,ϕ which determine the direction of the vector p_μ.

First it is straightforward to show that $|n,\pm\rangle,|n,\pm\rangle^*$ satisfy analogous relations to (3.10)–(3.13):

$$\not{p}\,|n,\pm\rangle_p=p\,|n\pm1,\mp\rangle^*_p, \tag{3.10′}$$
$$\not{p}\,|n,\pm\rangle^*_p=p\,|n\pm1,\mp\rangle_p, \tag{3.11′}$$
$$\not{\partial}\,|n,\pm\rangle_p=\left[\frac{d}{dp}\mp\frac{(n+1\mp1)}{p}\right]|n\pm1,\mp\rangle^*_p, \tag{3.12′}$$
$$\not{\partial}\,|n,\pm\rangle^*_p=\left[\frac{d}{dp}\mp\frac{(n+1\mp1)}{p}\right]|n\pm1,\mp\rangle_p, \tag{3.13′}$$

where $\not{p}=p_\mu\gamma_\mu$, $\not{\partial}=\gamma_\mu\partial_\mu$, and

$$\gamma_0=\begin{bmatrix}I&0\\0&-I\end{bmatrix}, \quad \boldsymbol{\gamma}=\begin{bmatrix}0&i\boldsymbol{\sigma}\\-i\boldsymbol{\sigma}&0\end{bmatrix}.$$

Now using the fact that, $Z_{n\pm}$, $\tilde{Z}_{n\pm}$ are bases for the SO(4) irreducible spaces $((n\pm1)/2,n/2)$ and $(n/2,(n\pm1)/2)$ respectively,[11] we can easily obtain the action of the Casimir operators of the SO(4) group, in the representation (3.7)–(3.8) on the $|n,\pm\rangle$, $|n,\pm\rangle^*$, and we find that they belong to the same representations that is to the $((n\pm1)/2,n/2)$, $(n/2,(n\pm1)/2)$, the only difference lying in the number of components.

Finally going to a polar coordinate system of the five dimensional Euclidean space,

$$\eta_5=\eta\cos\omega, \quad \eta_\mu=\eta\sin\omega\hat{\eta}_\mu, \quad \mu=1,2,3,4, \tag{3.16}$$

where $\hat{\eta}_\mu$ is a unit 4-vector with polar angles (θ_2,θ_1,ϕ), and defining the functions

$$|N,n,\pm,j,\mu\rangle_\eta \equiv \Omega_{N,n}(\omega)\cdot|n,\pm,j,\mu\rangle_{\hat\eta}, \tag{3.17}$$

$$|N,n,\pm,j,\mu\rangle^*_\eta \equiv \Omega_{N,n}(\omega)\ |n,\pm,j,\mu\rangle^*_{\hat\eta}, \tag{3.18}$$

where

$$\Omega_{N,n}(\omega)=\left[\frac{(N+3/2)(N-n)!}{(N+n+2)!}\right]^{n/2}\frac{P^{n+1}_{N+1}(\cos\omega)}{\sin\omega}, \tag{3.19}$$

we solve the eigenvalue problem for the first Casimir operator C_1 of the SO(5) group for $e=N+\frac{3}{2}$, $f=\frac{3}{2}$ [(3.5)]. In the representation (3.7)—(3.8) the C_1 operator is

$$C_1=C_{1,0}+I_{1,0}+(i/2)M^{5\mu}_{0\gamma}(\gamma^5\gamma^\mu-\gamma^\mu\gamma^5)+1, \tag{3.20}$$

where

$$C_{1,0}=\tfrac{1}{2}M^{ab}_{0\gamma}M^{ab}_{0\gamma},\ I_1=\tfrac{1}{2}M^{\mu\nu}M^{\mu\nu},\ I_{1,0}=\tfrac{1}{2}M^{\mu\nu}_{0\gamma}M^{\mu\nu}_{0\gamma},$$

and $a,b=1,2,3,4,5$, $\mu,\nu=1,2,3,4$. If we denote by D the operator

$$D\equiv(i/2)M^{5\mu}_{0\gamma}(\gamma^5\gamma^\mu-\gamma^\mu\gamma^5),$$

we can find, using the relations (3.10′)—(3.13′), the action of the operator D on the functions (3.17)—(3.18):

$$D|N,n,+,j,\mu\rangle_\eta=-\omega_{N,n}|N,n+1,-,j,\mu\rangle^*_\eta, \tag{3.21}$$

$$D|N,n,-,j,\mu\rangle_\eta=\omega_{N,n-1}|N,n-1,+,j,\mu\rangle^*_\eta, \tag{3.22}$$

$$D|N,n,+,j,\mu\rangle^*_\eta=\omega_{N,n}|N,n+1,-,j,\mu\rangle_\eta, \tag{3.23}$$

$$D|N,n,-,j,\mu\rangle^*_\eta=-\omega_{N,n-1}|N,n-1,+,j,\mu\rangle_\eta, \tag{3.24}$$

where $\omega_{N,n}=[(N-n)(N+n+3)]^{1/2}$. Since the functions (3.17)—(3.18) are eigenfunctions of the operators $C_{1,0}$, I_1, $I_{1,0}$ we find that the required normalized eigenfunctions of C_1 for $(e=N+\frac{3}{2},\ f=\frac{3}{2})$ will be eigenfunctions of the C_2 too and will be of the following form:

$$|N,+,((n+1)/2,n/2),j,\mu\rangle_\eta \equiv C_{N,n}|N,n,+,j,\mu\rangle_\eta+D_{N,n}|N,n+1,-,j,\mu\rangle^*_\eta, \tag{3.25}$$

$$|N+1,-,((n+1)/2,n/2),j,\mu\rangle_\eta \equiv D_{N+1,n}|N+1,n,+,j,\mu\rangle_\eta-C_{N+1,n}|N+1,n+1,-,j,\mu\rangle^*_\eta, \tag{3.26}$$

$$|N,+,(n/2,(n+1)/2),j,\mu\rangle_\eta \equiv D_{N,n}|N,n+1,-,j,\mu\rangle_\eta-C_{N,n}|N,n,+,j,\mu\rangle^*_\eta, \tag{3.27}$$

$$|N+1,-,(n/2,(n+1)/2),j,\mu\rangle_\eta \equiv C_{N+1,n}|N+1,n+1,-,j,\mu\rangle_\eta+D_{N+1,n}|N+1,n,+,j,\mu\rangle^*_\eta, \tag{3.28}$$

where

$$C_{N,n}=\left(\frac{N+n+3}{2N+3}\right)^{1/2},\quad D_{N,n}=\left(\frac{N-n}{2N+3}\right)^{1/2}$$

and the indices run as follows:

$$n=0,1,2,\ldots,N,\quad j=0,1,2,\ldots,n,$$
$$|\mu|=0,1,2,\ldots,j. \tag{3.29}$$

In the Appendix we show that the constructed $(e=N+\frac{3}{2},\ f=\frac{3}{2})$ SO(5) spinor spherical harmonics satisfy the following useful relations:

$$\hat\eta|N,+\rangle_\eta=-|N+1,-\rangle_\eta, \tag{3.30}$$

$$\hat\eta|N+1,-\rangle_\eta=-|N,+\rangle_\eta, \tag{3.31}$$

$$\gamma^5|N,+,((n+1)/2,n/2)\rangle_\eta =-[\rho_{N,n}|N,+,((n+1)/2,n/2)\rangle_\eta +\sigma_{N,n}|N,-,((n+1)/2,n/2)\rangle_\eta], \tag{3.32}$$

$$\gamma^5|N+1,-,((n+1)/2,n/2)\rangle_\eta =-[\sigma_{N+1,n}|N+1,+,((n+1)/2,n/2)\rangle_\eta -\rho_{N+1,n}|N+1,-,((n+1)/2,n/2)\rangle_\eta], \tag{3.33}$$

$$\gamma^5|N,+,(n/2,(n+1)/2)\rangle_\eta =-[-\rho_{N,n}|N,+,(n/2,n+1)/2)\rangle_\eta +\sigma_{N,n}|N,-,(n/2,(n+1)/2)\rangle_\eta], \tag{3.34}$$

$$\gamma^5|N+1,-,(n/2,(n+1)/2)\rangle_\eta =-[\sigma_{N+1,n}|N+1,+,(n/2,(n+1)/2)\rangle_\eta +\rho_{N+1,n}|N+1,-,(n/2,(n+1)/2)\rangle_\eta], \tag{3.35}$$

$$\int d\Omega_5(\xi)|N,\pm\rangle_\xi/(1-\xi\circ n)=[8\pi^2/(N+1)(N+2)]|N,\pm\rangle_\eta, \tag{3.36}$$

where

$$\hat\eta=\eta^a\gamma^a,\quad a=1,2,3,4,5,\quad \gamma^5=\gamma^1\gamma^2\gamma^3\gamma^4=\begin{bmatrix}0 & I\\ I & 0\end{bmatrix},$$

and

$$\rho_{N,n}=(2n+3)/(2N+3),\quad \sigma_{N,n}=2[(N-n)(N+n+3)]^{1/2}/2N+3. \tag{3.37}$$

In the above relations we have omitted SO(3) or SO(4) indices where they were unnecessary.

IV. SOLUTION OF THE B-S EQUATION AT $E_B = 0$

In this section we solve the B—S equation in the form (2.6). Considering Eq. (2.6) and the relations (3.30)—(3.36), we are sure, first of all, that the known SO(5) symmetry of the Wick—Cutkosky model for $E_B=0$ here is absent, but it is obvious that there is SO(4) symmetry. For this reason the bound state amplitude must be of the form

$$Y(\eta)=\sum_{N\geq N_0}[X_{N,n}|N,+,((n+1)/2,n/2)\rangle_\eta +Y_{N,n}|N+1,-,((n+1)/2,n/2)\rangle_\eta] \tag{4.1}$$

or of the form

$$\tilde Y(\eta)=\sum_{N\geq N_0}[\tilde X_{N,n}|N,+,(n/2,(n+1)/2)\rangle_\eta +\tilde Y_{N,n}|N+1,-,(n/2,(n+1)/2)\rangle_\eta], \tag{4.2}$$

where $N_0\geq n$.

Introducing the first form into Eq. (2.6) and using the relations (3.30)—(3.36), we find the recursion relations

$$Y_{N_0,n}=0,\quad \begin{bmatrix}X_{N+1,n}\\ Y_{N+1,n}\end{bmatrix}=\frac{1}{\sigma_{N+1,n}}\begin{bmatrix}\frac{1}{1-a_{N+1}} & \rho_{N+1,n}\\ \rho_{N+1,n} & 1-a_{N+1}\end{bmatrix}\begin{bmatrix}X_{N,n}\\ Y_{N,n}\end{bmatrix}, \tag{4.3}$$

where

$$N\geq N_0,\quad a_N=(N_0+1)(N_0+2)/(N+1)(N+2), \tag{4.4}$$

and $\rho_{N,n}$, $\sigma_{N,n}$ are given in (3.37).

The solution of the recursion relation (4.3) is

$$\begin{bmatrix} X_{N+1,n} \\ Y_{N+1,n} \end{bmatrix} = X_{N_0,n} \left(\prod_{k=N_0}^{N} \frac{1}{\sigma_{k+1,n}} \begin{bmatrix} 1/(1-a_{k+1}) & \rho_{k+1,n} \\ \rho_{k+1,n} & 1-a_{k+1} \end{bmatrix} \right) \times \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad N \geq N_0. \tag{4.5}$$

For the spectrum, if $X_{N_0,n} \neq 0$, we obtain

$$a = \tfrac{1}{2}(N_0+1)(N_0+2), \quad N_0 = 0,1,2,3,\cdots. \tag{4.6}$$

If we substitute the second form (4.2), we find the same spectrum, but now the solution of the corresponding recursion relation is

$$\begin{bmatrix} \tilde{X}_{N+1,n} \\ \tilde{Y}_{N+1,n} \end{bmatrix} = \tilde{X}_{N_0,n} \left(\prod_{k=N_0}^{N} \frac{1}{\sigma_{k+1,n}} \begin{bmatrix} \frac{1}{1-a_{k+1}} - \rho_{k+1,n} \\ -\rho_{k+1,n} \quad 1-a_{k+1} \end{bmatrix} \right) \times \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad N \geq N_0. \tag{4.7}$$

V. CONCLUSIONS

The use of the SO(5) spinor spherical harmonics, which are bases for the class $f=\frac{3}{2}$, $e=N+\frac{3}{2}$, $N=0,1,\cdots$, of the unitary irreducible representations of the SO(5) group, has given us in a natural way the spectrum of the coupling constant for $E_B=0$ and the bound state amplitudes.

The spectrum is SO(5) degenerate and coincides with that of the Wick–Cutkosky model, but the SO(5) symmetry is broken for the bound state amplitudes. This is due to the presence of the spin in the Yukawa type interaction.

The normalization problem[13] at $E_B=0$ requires the knowledge of the coupling constant dependence from E_B in a neighborhood of the point $E_B=0$. In the framework of our treatment, at least, this is rather cumbersome to obtain because it requires diagonalization of perturbational terms which break even the existing at $E_B=0$ SO(4) symmetry of the interaction.

ACKNOWLEDGMENTS

I am grateful to Professor Fokion Hadjioannou for suggesting the problem and for his encouraging guidance and advice during the progress of this work. I also express my thanks to Dr. N. Antoniou at N.R.C. "Democritus," for his interest for the realization of this paper. I wish to express my gratitude also to the Nation al Hellenic Research Foundation for the grant which made possible this work.

APPENDIX

In this appendix we prove the relations (3.21)–(3.24) and (3.30)–(3.36). First from the representation (3.8) of the operators $M_{0r}^{5\mu}$ we find that the operator D has the form

$$D = -\gamma^5 (n_5\, {}_4\partial\!\!\!/ - {}_4\eta\!\!\!/\,\partial_5), \tag{A1}$$

where ${}_4\partial\!\!\!/ = \gamma_\mu \partial/\partial\eta_\mu$, ${}_4\eta\!\!\!/ = \gamma_\mu \eta_\mu$, $\mu=1,2,3,4$. Using the polar coordinate system (3.16) and the relations (3.10′)–(3.13′), we find

$$D|N,n,\pm\rangle_\eta = \left[\sin\omega \frac{\partial}{\partial\cos\omega} \pm (n\mp 1+1)\cot\omega\right]\Omega_{N,n}(\omega)\,|n\pm 1,\mp\rangle^*, \tag{A2}$$

$$D|N,n,\pm\rangle_\eta^* = -\left[\sin\omega \frac{\partial}{\partial\cos\omega} \pm (n\mp 1+1)\cot\omega\right]\Omega_{N,n}(\omega)\,|n\pm 1,\mp\rangle. \tag{A3}$$

The definition of $\Omega_{N,n}(\omega)$, (3.19), and the recurrence relations of the Legendre functions (3.8), (11), (13), (14), (19) of p. 161 of Ref. 14 permit to us to extract the required relations (3.21)–(3.24). For the relations (3.30)–(3.31) we have to observe that

$$\gamma_5 |n,\pm\rangle = -|n,\pm\rangle, \tag{A4}$$

$$\gamma_5 |n,\pm\rangle^* = |n,\pm\rangle^* \tag{A5}$$

and that, for $n_a\eta_a=1$, $a=1,2,3,4,5$,

$$\eta\!\!\!/\,|N,\pm\rangle = [\cos\omega\gamma_5 + \sin\omega(\hat{\eta}_\mu\gamma_\mu)]\cdot|N,\pm\rangle, \quad \mu=1,2,3,4, \tag{A6}$$

where $\hat{\eta}_\mu$ is the unit four-vector in the relation (3.16). Using the relations (A4), (A5), (3.10′)–(3.13′), (3.25)–(3.28), and the recurrence relations (3.8) (12), (13), (14) of p. 161 in Ref. 14, we find the truth of (3.30) and (3.31).

For the relations (3.32)–(3.35), we solve the relations (3.25)–(3.26) with respect to $|N,n,\pm\rangle$, $|N,n,\pm\rangle^*$, and using (A4)–(A5), we substitute them in the action of γ^5 on the $|N,\pm\rangle$'s.

For the relations (3.36) we observe that[15]

$$\int d\Omega_s(\xi) Y_{Nnlm}(\hat{\xi})/(1-\xi\cdot n) = [8\pi^2/(N+1)(N+2)]\, Y_{Nnlm}(\hat{\eta}), \tag{A7}$$

where $Y_{Nnlm}(\hat{\xi})$ are the five-dimensional spherical harmonics. Because our $|N,\pm\rangle$ are columns of linear combinations of such functions for different n,l,m, we see immediately the truth of the relations (3.36).

*Work supported by National Hellenic Research Foundation.

[1] H. Bethe and E. Salpeter, Phys. Rev. 82, 309 (1951).
[2] N. Nakanishi, Suppl. Prog. Theor. Phys. No. 43 (1969).
[3] J. Ball and F. Zachariasen, Phys. Rev. 170, 1541 (1968).
[4] M. Ciafaloni and P. Menotti, Phys. Rev. 173, 1575 (1968).
[5] S. Drell and T. Lee, Phys. Rev. D 5, 1738 (1972).
[6] S. Singh, Phys. Rev. D 6, 1648 (1972).
[7] Y. Munakata and R. Sugano, Prog. Theor. Phys. 16, 532 (1956).
[8] V. Fock, Z. Phys. 98, 145 (1935).
[9] J. Kurijan, N. Mukunda, and E. Sudarsan, Commun. Math. Phys. 8, 204 (1968).
[10] G. C. Wick, Phys. Rev. 96, 1124 (1954).
[11] K. Rothe, Phys. Rev. 170, 1548 (1968).
[12] Ref. 11, p. 1554.
[13] See for instance, Ref. 2, p. 9.
[14] Erdelyi, Magnus, Oberhettinger, and Tricomi, *Higher Transcendental Functions* (McGraw-Hill, New York, 1953), Vol. I.
[15] R. L. Cutkosky, Phys. Rev. 96, 1135 (1954), Appendix A, p. 1140.

Asymptotic solutions and conservation laws for the nonlinear Schrödinger equation. I

Harvey Segur and Mark J. Ablowitz

Clarkson College of Technology, Potsdam, New York 13676
(Received 26 September 1975)

In the absence of solitons, the nonlinear Schrödinger equation has an asymptotic solution which decays in time as $t^{-1/2}$, and contains two arbitrary functions (in the amplitude and phase, respectively). For appropriate initial data, the amplitude function is uniquely determined in terms of the initial data by the conservation laws; the other function is undetermined. This method determines the leading two terms in each of the asymptotic expansions for the amplitude and phase, but no more. The method makes no direct use of th Marchenko integral equations.

INTRODUCTION

The "method of inverse scattering" has been applied to a number of special nonlinear partial differential equations (e.g., see Ref. 1), including the nonlinear Schrödinger equation[2,3]

$$iu_t + u_{xx} + 2\alpha |u|^2 = 0, \quad -\infty < x < \infty. \tag{1}$$

Here $\alpha = \pm 1$, and the two equations have significantly different solutions. If $\alpha = +1$, the equation has an N-soliton solution,[2] and the general solution consists of solitons (and bound states) traveling in a background of dispersive, decaying oscillations. As shown in Ref. 1, no permanent waves can arise if

$$\int_{-\infty}^{\infty} |u(x,0)|\, dx < 0.904, \tag{2}$$

and the solution that evolves from such initial data consists entirely of decaying oscillations. If $\alpha = -1$, any initial data that vanish as $|x| \to \infty$ evolve into decaying oscillations. In either case, the oscillations are associated with the continuous spectrum.

The detailed asymptotic structure of the oscillations was examined for the *KdV* equation in Ref. 4, and for the nonlinear Schrödinger equation in Ref. 5. In the former paper, we have found an error in the evaluation of certain integrals. We shall discuss this later. In both of the above papers, attention was restricted to the case of no solitons and the authors examined the Marchenko integral equations in the limit $t \to \infty$.

In this and the companion paper (I and II), we determine the asymptotic solution to (1) by an alternative method which uses the conservation laws rather than the integral equations. In the simplest case of no solitons (I), this method reproduces (in a slightly stronger form) the results obtained in Ref. 5. If $\alpha = -1$, this is the general asymptotic solution. If $\alpha = +1$, this is the asymptotic state only if the initial data are "small," in the sense of (2). Without such a restriction, the asymptotic state contains both solitons and decaying oscillations, and is discussed in the following paper.[6]

II. ASYMPTOTIC SOLUTION

There is a closed form similarity solution[7] to (1),

$$u(x,t) = t^{-1/2} A \exp\left\{it\left[\frac{1}{4}\left(\frac{x}{t}\right)^2 + 2\alpha A^2 \frac{\ln t}{t} + \frac{\phi}{t}\right]\right\}.$$

Guided by this solution, we seek another solution to (1) in which A, ϕ are slowly varying. The appropriate expansion turns out to be

$$u(x,t) = t^{-1/2} R \exp(it\theta),$$
$$R\left(\frac{x}{t}, t\right) = f\left(\frac{x}{t}\right) + \sum_{n=1}^{\infty} \sum_{k=0}^{n} \frac{(\ln t)^k}{t^n} f_{nk}\left(\frac{x}{t}\right), \tag{3}$$
$$\theta\left(\frac{x}{t}, t\right) = \frac{1}{4}\left(\frac{x}{t}\right)^2 + \sum_{n=1}^{\infty} \sum_{k=0}^{n} \frac{(\ln t)^k}{t^n} \theta_{nk}\left(\frac{x}{t}\right),$$

where R, θ are both real-valued functions. By direct substitution into (1), one finds that

$$\begin{aligned}
&f(x,t) \text{ arbitrary}, && \theta_{11} = 2\alpha f^2, \\
&f_{11} = 4\alpha f[3(f')^2 + ff''], && \theta_{10} = g(x/t), \text{ arbitrary} \\
&f_{10} = fg'' + 2g'f' && \theta_{22} = 16(ff')^2, \\
&\quad + 4\alpha f[3(f')^2 + ff''], && \theta_{21} = 2f[f'g' - 8(f(f')^2 + f^2 f'')], \\
&\vdots && \vdots
\end{aligned} \tag{4}$$

The expansion in (3) can be carried to any desired order in n. All the subsequent coefficients in the expansion can be found explicitly in terms of two arbitrary functions, $f(x/t)$ and $g(x/t)$. The two functions are unrestricted by Eq. (1), and are expected to be fixed by the initial data. In the subsequent analysis, we *assume* that the solution of (1) evolves from appropriate initial data into forms (3) and (4). We show that the conservation laws uniquely determine f, but place no restrictions on g. The simplectic form in the Hamiltonian formulation of this problem suggests what g might be, but this conjecture cannot be proved by this approach.

III. CONSERVATION LAWS

Zakharov and Faddeev[8] noted that for the *KdV* equation, the infinite set of conserved quantities can be identified with certain moments of the scattering data. This identification has since been shown to be a general property of inverse scattering problems,[9-12] and is summarized below. The key ingredient that is added here is to observe that if the conservation laws are written in terms of the asymptotic solution, (3) and (4), then this infinite sequence of equations uniquely determines $f(x,t)$, the asymptotic amplitude.

We review here certain aspects of the inverse scattering solution of (1); complete details are given in

Refs. 1 and 2. We assume throughout that the initial data $u(x,0)$ is infinitely differentiable and vanishes rapidly as $|x| \to \infty$. The associated linear eigenvalue problem is

$$\begin{aligned} v_{1_x} + i\zeta v_1 &= uv_2, \\ v_{2_x} - i\zeta v_2 &= -\alpha u^* v_1. \end{aligned} \tag{5}$$

Two linearly independent solutions (not complex conjugates) are defined by

$$\left.\begin{aligned} \phi(x,\zeta) &\to \binom{1}{0}\exp(-i\zeta x), \\ \bar{\phi}(x,\zeta) &\to \binom{0}{-1}\exp(i\zeta x), \end{aligned}\right\} x \to -\infty.$$

The scattering data are then defined by

$$\left.\begin{aligned} \phi &\to \begin{pmatrix} a(\zeta)\exp(-i\zeta x) \\ b(\zeta)\exp(i\zeta x) \end{pmatrix}, \\ \bar{\phi} &\to \begin{pmatrix} \bar{b}(\zeta)\exp(-i\zeta x) \\ -\bar{a}(\zeta)\exp(i\zeta x) \end{pmatrix}, \end{aligned}\right\} x \to +\infty. \tag{6}$$

The time dependence of the scattering data can be written

$$\begin{aligned} &\frac{\partial}{\partial t}(\ln a) = 0 = \frac{\partial}{\partial t}\ln\left(1 - \alpha\left|\frac{b}{a}\right|^2\right), \\ &\frac{\partial}{\partial t}(\arg b(\xi)) = 4\xi^2, \end{aligned} \tag{7}$$

by making use of the identities (for real ξ)

$$a\bar{a} + b\bar{b} = 1, \quad \bar{b} = -\alpha b^*, \quad \bar{a} = a^*.$$

In the absence of any discrete spectrum, $\ln a(\zeta)$ ($\ln\bar{a}(\zeta)$) is analytic in the upper (lower) half-plane, vanishes as $|\zeta| \to \infty$ there, and is a constant of the motion. An asymptotic expansion of $\ln a$ for large ζ gives the infinite set of conserved densities[2]

$$\ln a(\zeta) = \sum_1^\infty (2i\zeta)^{-n} C_n, \tag{8a}$$

where

$$C_n = \int_{-\infty}^{\infty} d_n(x)\,dx, \tag{8b}$$

$$d_1(x) = \alpha\,|u(x)|^2; \tag{8c}$$

and for $n \geq 1$,

$$d_{n+1}(x) = u\frac{d}{dx}\left(\frac{d_n}{u}\right) + \sum_1^{n-1} d_k d_{n-k}. \tag{8d}$$

As examples, the next two conserved quantities are

$$\begin{aligned} C_2 &= \int_{-\infty}^{\infty} (\alpha u u_x^*)\,dx, \\ C_3 &= \int_{-\infty}^{\infty} (\alpha u u_{xx}^* + |u|^4)\,dx. \end{aligned} \tag{9}$$

A second asymptotic expansion of $\ln a$ yields the "trace formulas." Based on the analytic properties of $\ln a$ and Cauchy's integral theorem, it is not difficult to show that

$$\ln a = +\frac{1}{2\pi i}\sum_1^\infty \zeta^{-n} \times \int_{-\infty}^{\infty} \xi^{n-1}\ln\left(1 - \alpha\left|\frac{b}{a}(\xi)\right|^2\right) d\xi. \tag{10}$$

It follows from (8) that $\ln[1 - \alpha\,|(b/a)|^2]$ is transcendentally small as $|\xi| \to \infty$, so that all the integrals in (10) converge. Equating coefficients in (8a) and (10), one obtains for $n = 1, 2, \cdots$,

$$(2i)^{-n} C_n = (2\pi i)^{-1}\int_{-\infty}^{\infty} \xi^{n-1}\ln\left(1 - \alpha\left|\frac{b}{a}\right|^2\right) d\xi. \tag{11}$$

These identities were first obtained (for *KdV*) in Ref. 7.

The conserved quantities C_n, when written in terms of the asymptotic solution, (3) and (4), take on a simple form. The first three are

$$\begin{aligned} C_1 &= \alpha\int_{-\infty}^{\infty} f^2\left(\frac{x}{t}\right) d\left(\frac{x}{t}\right), \\ C_2 &= \alpha\int_{-\infty}^{\infty} \left(\frac{x}{2it}\right) f^2\left(\frac{x}{t}\right) d\left(\frac{x}{t}\right), \\ C_3 &= \alpha\int_{-\infty}^{\infty} \left(\frac{x}{2it}\right)^2 f^2\left(\frac{x}{t}\right) d\left(\frac{x}{t}\right). \end{aligned} \tag{12}$$

We clain that every conserved density has the form

$$d_n = \alpha\left(\frac{x}{2it}\right)^{n-1} f^2\left(\frac{x}{t}\right) + O\left(\frac{\ln t}{t}\right). \tag{13}$$

The proof is by induction. Certainly d_1, d_2, d_3 obey (13). If the first n conserved densities obey (13), then it follows from (8d) that d_{n+1} does as well. Consequently, the identities (11) become, for all $n \geq 1$,

$$\alpha\int_{-\infty}^{\infty} \left(\frac{-x}{4t}\right)^{n-1} f^2\left(\frac{x}{t}\right) d\left(\frac{x}{t}\right) = \frac{1}{\pi}\int_{-\infty}^{\infty} \xi^{n-1}\ln\left(1 - \alpha\left|\frac{b}{a}(\xi)\right|^2\right) d\xi. \tag{14}$$

This infinite set of moment equations can be satisfied only if

$$\begin{aligned} &\xi = -x/4t, \\ &f^2(-4\xi) = \frac{\alpha}{4\pi}\ln\left(1 - \alpha\left|\frac{b}{a}(\xi)\right|^2\right). \end{aligned} \tag{15}$$

Equations (15), along with (3) and (4), completely relate the leading two terms in the asymptotic solution of (1) to the initial data. Combining these results, we obtain

$$u(x,t) = t^{-1/2}R[(x/t),t]\exp\{it\theta[(x/t),t]\},$$

where

$$\begin{aligned} \theta &= \frac{1}{4}\left(\frac{x}{t}\right)^2 + 2\alpha f^2\left(\frac{x}{t}\right)\frac{\ln t}{t} + O\left(\frac{1}{t}\right), \\ R &= f\left(\frac{x}{t}\right) + 4\alpha f\left(\frac{x}{t}\right)\left[3\left(f'\left(\frac{x}{t}\right)\right)^2 + f\left(\frac{x}{t}\right)f''\left(\frac{x}{t}\right)\right]\frac{\ln t}{t} + O\left(\frac{1}{t}\right), \end{aligned} \tag{16}$$

and

$$f^2\left(\frac{x}{t}\right) = \frac{\alpha}{4\pi}\ln\left[1 - \alpha\left|\frac{b}{a}\left(-\frac{x}{4t}\right)\right|^2\right].$$

This is the main result in this paper. In Ref. 5, the first terms of these series [(16b) and (16c)] were obtained, but (16d) was only obtained in an integral sense. Those results, combined with Eqs. (3) and (4), are exactly equivalent to (16).

IV. INVERSE SCATTERING AS A CANONICAL TRANSFORMATION

In the Hamiltonian formulation of inverse scattering,[8-12] $\ln[1-\alpha|(b/a)|^2]$ is shown to be an action variable, and therefore a constant of the motion. Equation (15b) suggests that the asymptotic amplitude also can be considered an action variable. We now pursue this line of reasoning, which suggests (but does not uniquely determine) the relation between $g(x/t)$ and the initial data. With

$$u=\rho e^{i\phi}, \tag{17}$$

three canonical formulations of (1), or equivalently of (7), are

$$P_1=u^*,\quad Q_1=iu,\quad H_1=\int_{-\infty}^{\infty}(|u_x|^2-\alpha|u|^4)\,dx, \tag{18}$$

$$P=\rho^2,\quad Q=-\phi,\quad H=H_1=\int_{-\infty}^{\infty}(\rho_x^2+\rho^2\phi_x^2-\alpha\rho^4)\,dx, \tag{19}$$

$$p=\frac{\alpha}{\pi}\ln\left(1-\alpha\left|\frac{b}{a}\right|^2\right),\quad q=\arg b, \tag{20}$$

$$K=\frac{4\alpha}{\pi}\int_{-\infty}^{\infty}\xi^2\ln\left(1-\alpha\left|\frac{b}{a}\right|^2\right)d\xi.$$

The transformation between (18) and (20) was shown to be canonical in Ref. 12, and one easily verifies that the Poisson brackets between (18) and (19) are invariant. It follows that all these transformations are canonical, and preserve the simplectic form:

$$\int_{-\infty}^{\infty}(dp\wedge dq)\,dx\equiv\int_{-\infty}^{\infty}(\delta_1 p\,\delta_2 q-\delta_2 p\,\delta_1 q)\,dx, \tag{21}$$

where δ_1, δ_2 refer to independent variations. For (20), this form is

$$\int_{-\infty}^{\infty}d\left[\frac{2\alpha}{\pi}\ln\left(1-\alpha\left|\frac{b}{a}(\xi)\right|^2\right)\right]\wedge d(\arg b(\xi))\,d\xi.$$

For (19), using (3) and (4), we obtain an infinite sequence of forms, asymptotically ordered in t. Within these forms, there are only three independent variations, involving $f(x/t)$, $g(x/t)$, and t. The first nontrivial form in this sequence is

$$\int_{-\infty}^{\infty}d\left(f^2\left(\frac{x}{t}\right)\right)\wedge d\left(-g\left(\frac{x}{t}\right)\right)d\left(\frac{x}{t}\right),$$

and all others vanish asymptotically. Equating these two forms and using (15), one obtains

$$\int_{-\infty}^{\infty}d\left[\ln\left(1-\alpha\left|\frac{b}{a}\right|^2\right)\right]\wedge d(g(-4\xi)+\arg b(\xi))\,d\xi=0. \tag{22}$$

It is tempting to conclude that

$$g_0(-4\xi)=-\arg b(\xi)=\arg b^*(\xi), \tag{23}$$

and this is a possible solution. Another possibility is

$$g_1(-4\xi)=\arg b^*(\xi)+\tfrac{3}{4}\pi, \tag{24}$$

which will be convenient below. In fact, (22) allows us to add to $g(x/t)$ any function of $f(x/t)$. Practically, this means that the first two terms in the expansion (3) can be determined from the initial data, but that no higher terms can be obtained by this approach.

V. DISCUSSION OF RESULTS

Let us first relate the results obtained here to the solution of the linearized problem. In the linear limit, where

$$u(x,0)=\epsilon q(x),\quad \epsilon\ll 1\quad\text{and}\quad\int_{-\infty}^{\infty}|q|\,dx=O(1),$$

one can show that[1]

$$a(\xi)=1+O(\epsilon^2),$$

$$b^*(\xi)=-\epsilon\alpha\int_{-\infty}^{\infty}q(x)\exp(-2i\xi x)\,dx+O(\epsilon^3) \tag{25}$$

$$=-\epsilon\alpha\,\hat{q}(-2\xi)+O(\epsilon^3).$$

Choosing (24), and expanding in power of $|b/a|$, our result can be written

$$u(x,t)=-(4\pi t)^{-1/2}\frac{b^*}{a^*}\left(-\frac{x}{4t}\right)\left(1-\frac{\alpha}{4}\left|\frac{b}{a}\right|^2+\cdots\right)$$
$$\times\exp\left\{i\left[\frac{t}{4}\left(\frac{x}{t}\right)^2-\frac{\pi}{4}+\frac{\alpha}{2\pi}\left|\frac{b}{a}\right|^2\ln t+\cdots\right]\right\}. \tag{26}$$

For comparison, one could linearize (1), solve the linear equation by Fourier transforms, and evaluate the results asymptotically by stationary phase. The result, after using (25), is

$$u(x,t)\sim-(4\pi t)^{-1/2}\frac{b^*}{a^*}\left(-\frac{x}{4t}\right)\exp\left\{i\left[\frac{t}{4}\left(\frac{x}{t}\right)^2-\frac{\pi}{4}\right]\right\}. \tag{27}$$

A comparison of (26) and (27) yields several conclusions. First, the justification for the choice (24) is that it is consistent with the linear result. Given this choice, the linearized limit of the nonlinear solution reproduces the linear solution exactly. Second, the appropriate small parameter for this linearization is $\max(|b/a|)$, or

$$\int_{-\infty}^{\infty}|u(x,0)|\,dx\ll 1, \tag{28}$$

rather than any local amplitude. Third, (15a) can be interpreted as defining the group velocity, a concept which is ordinarily associated with linear problems, but which appears naturally in the solution of this nonlinear problem. Thus, the "nonlinear" part of this problem consists of mapping the initial data into transform space. The subsequent evolution of the oscillatory waves is essentially linear.

Next, we wish to comment upon the asymptotic evaluation of the Marchenko integral equations in Ref. 4. We have found that there are errors in the stationary phase calculation of certain multiple integrals in the oscillatory region of *KdV* (there are also regions of exponential decay, and similarity). Due to these errors, the results in Ref. 4 are incorrect in this region. In the nonlinear Schrödinger equation, the asymptotic state (without solitons present) is decaying oscillations on $-\infty<x<\infty$. The methods in Ref. 4 apply formally, but the correct stationary phase evaluation of the resulting integrals makes the calculation extremely complicated.

A typical integral one has to compute in the oscillatory region is

$$\int_0^\infty \int_0^\infty q(x,y)\exp(ixyt)\,dx\,dy \underset{t\to\infty}{\sim} iq(0,0)\frac{\ln t}{t}. \tag{28}$$

Due to the fact that the stationary point lies at $x=y=0$, and to the form of the rapid phase, we have a logarithmically larger contribution than one usually obtains by conventional multidimensional stationary phase.[13] In Ref. 4, the leading terms were found by multiplying the standard stationary phase result by a factor depending on the dimension of the corner. This is true only if the rapid phase in the integrand is a "center" (see Ref. 13), and is oscillatory on the boundary.

In order to correct the results in Ref. 4, one must keep the $\ln t/t$ and $1/t$ terms in each of the integrals. Whereas the procedures suggest an asymptotic solution such as (3), the detailed analysis is quite complicated.

In summary, when no solitons exist, the asymptotic solution of (1) has a linearlike structure: waves propagate with their (linear) group velocity; the decay rate ($t^{-1/2}$) can be attributed entirely to (linear) frequency dispersion. The dominant nonlinear effects comes in mapping the initial data into transform space. This asymptotic solution can be obtained either by solving the integral equations approximately or, as demonstrated here, by utilizing the infinite set of conservation laws.

In terms of the method, the asymptotic ($t\to\infty$) form of the solution contains two arbitrary functions, related to the amplitude and phase. The conservation laws determine the leading contribution to the amplitude. The leading two terms in the phase are found explicitly, but the third term, which is 0(1), is not.

It is well known that the conservation laws are related to the action variables (cf. Ref. 8), which are "half" the information needed to solve the equation. It is remarkable that it is exactly *this* half of the information which determines the asymptotic amplitude. The other half, the angle variables, give some information about the asymptotic phase, but no explicit formulas.

ACKNOWELDGMENTS

We are grateful for helpful conversations with L. D. Faddeev, D. J. Kaup, and A. C. Newell. This work was supported by NSF Grant Nos. DES75-06537 and MPS75-07568.

[1]M. J. Ablowitz, D. J. Kaup, A. C. Newell, and H. Segur, Stud. Appl. Math. **53**, 249 (1974).
[2]V. E. Zakharov and A. B. Shabat, Zh. Eksp. Teor. Fiz. **61**, 118 (1971) [Sov. Phys. JETP **34**, 62 (1972)].
[3]M. J. Ablowitz, D. J. Kaup, A. C. Newell, and H. Segur, Phys. Rev. Lett. **31**, 125 (1973).
[4]M. J. Ablowitz and A. C. Newell, J. Math. Phys. **14**, 1277 (1973).
[5]S. V. Manakov, Zh. Eksp. Teor. Fiz. **65**, 1392 (1973)[Sov. Phys. JETP **38**, 693 (1974)].
[6]H. Segur, following paper, J. Math. Phys. **17**, 714 (1976).
[7]D. J. Benney and A. C. Newell, J. Math. Phys. **46**, 133 (1967).
[8]V. E. Zakharov and L. D. Faddeev, Funct. Anal. Appl. **5**, 10 (1971).
[9]D. W. McLaughlin, J. Math. Phys. **16**, 96 (1975).
[10]H. Flaschka and A. C. Newell, *Dynamical Systems, Theory and Applications*, edited by J. Moser (Springer-Verlag, Berlin, 1975).
[11]D. J. Kaup, J. Math. Phys. (to be published).
[12]V. E. Zakharov and S. V. Manakov, Teor. Mat. Fiz. **19**, 332 (1974).
[13]N. Bleistein and R. A. Handelsman, J. Math. Anal. Appl. **27**, 434 (1969).

Asymptotic solutions and conservation laws for the nonlinear Schrödinger equation. II

Harvey Segur

Department of Mathematics, Clarkson College of Technology, Potsdam, New York 13676
(Received 26 September 1975)

We find the dominant asymptotic behavior of the solution of the nonlinear Schrödinger equation when there is one soliton and decaying oscillations. The solution behaves like the soliton near the soliton, and like the solution found in the preceding paper (I) elsewhere. The method of solution uses the conservation laws, rather than the integral equations.

I. INTRODUCTION

Asymptotic solutions of the nonlinear Schrödinger equation[1]

$$iu_t + u_{xx} + 2|u|^2 u = 0, \quad -\infty < x < \infty \tag{1}$$

have been found for several classes of initial data. The soliton solution associated with the eigenvalue $(\xi + i\eta)$ is

$$u(x,t) = 2\eta \exp[i\phi(x,t)]\,\mathrm{sech}\psi(x,t), \tag{2}$$

where

$$\phi = -2[\xi x + 2(\xi^2 - \eta^2)t] + \phi_0,$$
$$\psi = 2\eta(x + 4\xi t) + \psi_0.$$

More generally, N-soliton solutions and multisoliton bound states, both associated with purely discrete spectra in the related scattering problem, are discussed in Ref. 1. Alternatively, if the spectrum is purely continuous, the asymptotic solution consists of decaying oscillations, as found in Ref. 2 and in a preceding paper[3]:

$$u(x,t) = t^{-1/2} R[(x/t), t] \exp\{it\theta[(x/t), t]\}, \tag{3}$$

where

$$R\left(\frac{x}{t}, t\right) = \frac{1}{4\pi} \ln\left\{1 + \left|\frac{b}{a}\left(-\frac{x}{4t}\right)\right|^2\right\}^{1/2} + O\left(\frac{\ln t}{t}\right),$$
$$\theta\left(\frac{x}{t}, t\right) = \frac{1}{4}\left(\frac{x}{t}\right)^2 + O\left(\frac{\ln t}{t}\right).$$

Here (b/a) is related to the initial data through the associated scattering problem (see Ref. 3). Arbitrary initial data, of course, can generate both discrete and continuous spectra, and one expects the general asymptotic solution to be some combination of solutions like (2) and (3).

In this paper we derive the dominant asymptotic behavior of the solution of (1) which contains both a soliton and decaying oscillations. It will turn out that in this case the asymptotic solution has the form

$$u(x,t) = 2\eta \exp(i\phi)\,\mathrm{sech}\psi + t^{-1/2} R\left(\exp(it\theta)\frac{(\xi + x/4t + i\eta\tanh\psi)^2}{(\xi + x/4t)^2 + \eta^2}\right.$$
$$\left. + \exp(2i\phi - it\theta)\frac{\eta^2 \mathrm{sech}^2\psi}{(\xi + x/4t)^2 + \eta^2}\right) + \cdots, \tag{4}$$

where we have used the notation of (2) and (3). Thus, the solution behaves as in (2) near the soliton, and as in (3) away from the soliton. [The phase of the solution, which is not entirely specified by these formulas, may differ between (4) and the other two problems cited.]

The method employed here also can be applied to the case of N solitons plus decaying oscillations, but the calculations become unwieldy and have not been performed. However, if the N solitons travel with N different speeds, the generalization is immediate. The solitons separate after a short time, and the long-time evolution of each soliton is as described herein.

As in Ref. 3, the method used to obtain the solution uses the conservation laws, rather than the integral equations. The leading terms of an asymptotic solution of (1) contains decaying oscillations, whose amplitude is an arbitrary constant. This solution is valid locally, and there is a neighboring asymptotic solution in which the amplitude is an unknown, slowly varying function of (x/t). Finally, the trace formulas, when written in terms of this asymptotic solution, uniquely determine the unknown amplitude function, and thereby the dominant asymptotic behavior of the solution. As found in Ref. 3, the phase of the solution is not entirely determined by this approach.

II. ASYMPTOTIC SOLUTION

It is necessary to find a solution to (1) in which the leading term is given by (2) and the next term decays like $t^{-1/2}$. In order to do so, one can either perturb the soliton solution of the differential equation (1), or perturb the corresponding solution of the integral equations. Either method is legitimate; we describe the former.

Using the definitions in (2) and (3), we seek a solution of (1) in the form

$$u(x,t) = 2\eta \exp(i\phi)\,\mathrm{sech}\psi + t^{-1/2}(f(\psi)\exp(it\theta) + g(\psi)\exp(2i\phi - it\theta)) + \cdots. \tag{5}$$

Substituting into (1), one finds that f and g must satisfy a coupled set of linear ordinary differential equations:

$$f'' + 2i(\alpha/\eta)f' + 2\,\mathrm{sech}^2\psi(2f + g^*) = 0,$$
$$(g^*)'' + 2i(\alpha/\eta)(g^*)' - 2[1 + (\alpha/\eta)^2]g^* + 2\,\mathrm{sech}^2\psi(2g^* + f) = 0, \tag{6}$$

where

$$\alpha = \xi + x/4t = (1/t)[(\psi - \psi_0)/B\eta],$$

( )* denotes complex conjugate, and we treat (x/t), or equivalently ψ/t, as a constant in this multiple scales approach. Equations (6) have bounded solutions

$$f(\psi) = A(\alpha + i\eta \tanh\psi)^2, \quad g(\psi) = A^*\eta^2 \operatorname{sech}^2\psi, \tag{7}$$

where A, which determines the amplitude of the oscillations, is an arbitrary constant.

Thus, to leading order, an asymptotic solution of (1) is

$$\begin{aligned} u(x,t) = {} & 2\eta \exp(i\phi) \operatorname{sech}\psi \\ & + t^{-1/2}[A(\alpha + i\eta \tanh\psi)^2 \exp(it\theta) \\ & + A^*\eta^2 \operatorname{sech}^2\psi \exp(2i\phi - it\theta)] + \cdots . \end{aligned} \tag{8}$$

In order to carry this expansion procedure to higher orders, A is considered constant with respect to the "fast" variables $(\psi, \phi, t\theta)$, but can depend on the "slow" variable $(x/t$, or $\psi/t)$. The expansion generated in this way is not uniform in ψ, and we cannot write the nth term of the expansion. Fortunately, none of these terms are needed in the subsequent analysis. Henceforth, we assume that (8), with $A = A(x/t)$, is an asymptotic representation of the solution of (1) when there is one soliton. It remains to determine $A(x/t)$ from the initial data.

III. CONSERVATION LAWS

If the initial data for (1) is infinitely smooth and vanishes rapidly as $|x| \to \infty$, there are an infinite number of constants of the motion, the "conserved densities."[1] As in Ref. 3, the trace formulas relate these densities to the scattering data via asymptotic expansions (for large ζ) of $\ln a(\zeta)$. In the case being considered, the spectrum contains one discrete eigenvalue ($\zeta_0 = \xi + i\eta$, $\eta > 0$) and a continuous spectrum. The appropriate expansion of $\ln a(\zeta)$ is

$$\begin{aligned} \ln a(\zeta) = \sum_1^\infty \zeta^{-n} \Bigg[& \frac{(\zeta_0^*)^n - \zeta_0^n}{n} \\ & + \frac{1}{2\pi i} \int_{-\infty}^{\infty} k^{n-1} \ln\left(1 + \left|\frac{b}{a}(k)\right|^2\right) dk \Bigg]. \end{aligned} \tag{9}$$

Meanwhile, the expansion of $\ln a$ in terms of the conserved densities is unchanged from Ref. 3:

$$\ln a(\zeta) = \sum_1^\infty (2i\zeta)^{-n} C_n, \tag{10a}$$

where

$$C_n = \int_{-\infty}^{\infty} d_n(x)\, dx, \tag{10b}$$

$$d_1(x) = |u(x)|^2, \tag{10c}$$

and for $n \geq 1$

$$d_{n+1}(x) = u \frac{d}{dx}\left(\frac{d_n}{u}\right) + \sum_1^{n-1} d_k d_{n-k}. \tag{10d}$$

Equating coefficients in (9) and (10), one obtains for $n = 1, 2, 3, \ldots,$

$$\begin{aligned} (2i)^{-n} C_n = {} & \frac{(\zeta_0^*)^n - \zeta_0^n}{n} \\ & + (2\pi i)^{-1} \int_{-\infty}^{\infty} k^{n-1} \ln\left(1 + \left|\frac{b}{a}(k)\right|^2\right) dk. \end{aligned} \tag{11}$$

It there were no continuous spectrum, one could compute, from (2) and (10), the conserved densities associated entirely with the soliton. For example,

$$\begin{aligned} \tilde{c}_1(x) &= 4\eta^2 \operatorname{sech}^2\psi, \\ \tilde{c}_2(x) &= 8i\xi\eta^2 \operatorname{sech}^2\psi - 8\eta^3 \operatorname{sech}^2\psi \tanh\psi, \end{aligned} \tag{12}$$

and we denote by $\tilde{c}_n(x)$ the integrand for the nth conserved density associated entirely with the soliton. Moreover, the integral in (11) would vanish, and the densities themselves could also be computed simply from (10):

$$\int_{-\infty}^{\infty} \tilde{c}_n(x)\, dx = (2i)^n \left(\frac{(\zeta_0^*)^n - \zeta_0^n}{n}\right). \tag{13}$$

When there is a continuous spectrum, the conserved densities still take on a simple asymptotic form when written in terms of the asymptotic solution (8). Each $\tilde{c}_n(x)$ is now the leading term in the expansion of the corresponding $d_n(x)$, but in each such expansion there is an $O(t^{-1})$ term whose integral is $O(1)$. For example, if $n = 1$, it follows from (8) and (10c) that

$$\begin{aligned} d_1(x) = {} & \tilde{c}_1(x) + t^{-1/2}\{\cdots\} \operatorname{sech}\psi + t^{-1}|A(x/t)|^2 \\ & \times |\xi + x/4t + i\eta \tanh\psi|^4 + \cdots, \end{aligned} \tag{14}$$

where $\{\cdots\}$ denotes bounded quantities and $\tilde{c}_1(x)$ is defined in (12). Integration yields

$$\begin{aligned} C_1 = \int_{-\infty}^{\infty} d_1(x)\, dx = {} & 4\eta + \int_{-\infty}^{\infty} \left|A\left(\frac{x}{t}\right)\right|^2 \\ & \times \left[\left(\xi + \frac{x}{4t}\right)^2 + \eta^2\right]^2 d\left(\frac{x}{t}\right) + O(t^{-1/2}). \end{aligned} \tag{15}$$

We claim that every $d_n(x)$ has the form

$$\begin{aligned} d_n(x) = {} & \tilde{c}_n(x) + t^{-1/2}\{\cdots\} \operatorname{sech}\psi \\ & + t^{-1}(x/2it)^{n-1} |A(x/t)|^2 |\xi + x/4t + i\eta \tanh\psi|^4 + \cdots . \end{aligned} \tag{16}$$

The proof is by induction, as shown in the Appendix. The conserved densities are obtained by integration

$$\begin{aligned} C_n = {} & \int_{-\infty}^{\infty} \tilde{c}_n(x)\, dx + \int_{-\infty}^{\infty} \left(\frac{x}{2it}\right)^{n-1} \left|A\left(\frac{x}{t}\right)\right|^2 \\ & \times \left[\left(\xi + \frac{x}{4t}\right)^2 + \eta^2\right]^2 d\left(\frac{x}{t}\right) + O(t^{-1/2}). \end{aligned} \tag{17}$$

Combining (11), (13), (15), and (17), one obtains, for $n \geq 1$,

$$\begin{aligned} & \int_{-\infty}^{\infty} \left(-\frac{x}{4t}\right)^{n-1} \left|A\left(\frac{x}{t}\right)\right|^2 \left[\left(\xi + \frac{x}{4t}\right)^2 + \eta^2\right]^2 d\left(\frac{x}{t}\right) + O(t^{-1/2}) \\ & = \frac{1}{\pi} \int_{-\infty}^{\infty} k^{n-1} \ln\left[1 + \left|\frac{b}{a}(k)\right|^2\right] dk. \end{aligned} \tag{18}$$

These equations have a unique solution

$$k=-x/4t,$$
$$\left|A\left(\frac{x}{t}\right)\right|^2=\frac{1}{4\pi}\frac{\ln[1+|(b/a)(-x/4t)|^2]}{[(\xi+x/4t)^2+\eta^2]^2}. \tag{19}$$

Substituting this result back into (8), and absorbing the unknown phase of $A(x/t)$ into an $O(1/t)$ correction of θ, one obtains the asymptotic solution of (1) given in (4).

As in Ref. 3, Eq. (19a) can be interpreted as defining the group velocity of the decaying oscillations. Thus, the complete solution, Eq. (4), consists of two distinct and largely unrelated parts. First, the soliton is a permanent, local and essentially nonlinear wave. It cannot be obtained by any linearization procedure. Second, the oscillations propagate with their linear group velocity, and decay in amplitude because of their linear frequency dispersion. Away from the soliton, they propagate almost as if (1) were linear.

IV. GENERALIZATIONS AND LIMITATIONS

We conclude with two comments about the implications of these results about other problems. First, in similar problems[4,5] methods have been developed and applied in which arbitrary initial data are assumed to evolve into N solitons (only), and the conservation laws are then used to identify these N solitons. It is obvious from the results derived here that although the nonsoliton part of the solution eventually disappears, it is not true that its contribution to the conserved densities disappears. Consequently, these methods can only be employed in problems in which there is *a priori* knowledge that the solitons contain a large fraction of the total energy available in the initial data.

Second, as pointed out in Ref. 6, the linear eigenvalue problem associated with (1) allows a real discrete eigenvalue ($\zeta_0=\xi+i\eta$, $n=0$). In such a case the Marchenko integral equations, as derived in Ref. 1, become singular. However, the trace formulas used here also break down. If we simply take the limit $n\to 0$ in (4), the solution exhibits a logarithmic singularity near $x/t=-4\xi$. [*Note added in proof*: More recent work suggests that, in this singular case, the local decay rate changes from $(1/t)^{1/2}$ to $(\ln t/t)^{1/2}$. Details will be published later.] It is worth noting that this logarithmic singularity, which occurs here as a special case, appears to be a general feature of the oscillatory solution of the Korteweg-deVries equation.

This work was supported by NSF Grant Nos. DES75-06537 and MPS75-07568.

APPENDIX

We prove by induction that every $d_n(x)$ has the form given in (16). It was shown in (14) that $d_1(x)$ has this form. We assume that each $d_j(x)$ $(j=1,\ldots,n)$ is of this form, and show from (10d) that $d_{n+1}(x)$ is as well.

First, the contribution from the product is

$$\sum_1^{n-1}\tilde{c}_k\tilde{c}_{n-k}+t^{-1/2}\,\mathrm{sech}\psi\{\cdots\}+\cdots. \tag{A1}$$

Second, for the derivative in (10d), we use the notation

$$\lambda=(\xi+x/4t+i\eta\tanh\psi)^2,\quad \tilde{u}=2\eta\exp(i\phi)\,\mathrm{sech}\psi, \tag{A2}$$

so that

$$\frac{d_n(x)}{u(x)}=\frac{\tilde{c}_n(x)+t^{-1}(x/2it)^{n-1}|A|^2|\lambda|^2+t^{-1/2}\,\mathrm{sech}\psi\{\cdots\}+\cdots}{\tilde{u}+t^{-1/2}(A\lambda\exp(it\theta)+\cdots)+\cdots}.$$

One can show that

$$u\frac{d}{dx}\left(\frac{d_n(x)}{u}\right)=\frac{(\tilde{u})^2(d/dx)(\tilde{c}_n/\tilde{u})}{\tilde{u}+t^{-1/2}(A\lambda\exp(it\theta)+\cdots)+\cdots}+\frac{(|A|^2A/t^{3/2})(x/2it)^n|\lambda|^2\lambda\exp(it\theta)}{\tilde{u}+t^{-1/2}(A\lambda\exp(it\theta)+\cdots)+\cdots}+\cdots. \tag{A3}$$

The first term on the right-hand side is significant only near $\psi=0$. In this region, the leading contribution of the first term is

$$\tilde{u}\frac{d}{dx}\left(\frac{\tilde{c}_n}{\tilde{u}}\right), \tag{A4}$$

and the second term provides only a higher-order correction to this term. Far away from $\psi=0$, the first term vanishes, $\tilde{u}$ vanishes, and the second term becomes

$$t^{-1}|A|^2(x/2it)^n\{|\lambda|^2+\cdots\}. \tag{A5}$$

Collecting terms from (A1), (A4), and (A5), one obtains

$$\begin{aligned}d_{n+1}(x)&=\tilde{u}\frac{d}{dx}\left(\frac{\tilde{c}_n}{\tilde{u}}\right)+\sum_1^{n-1}\tilde{c}_k\tilde{c}_{n-k}\\&\quad+t^{-1}|A|^2\left(\frac{x}{2it}\right)^n|\lambda|^2+\cdots\\&=\tilde{c}_{n+1}(x)+t^{-1}|A|^2\left(\frac{x}{2it}\right)^n\\&\quad\times\left[\left(\xi+\frac{x}{4t}\right)^2+\eta^2\tanh^2\psi\right]^2+\cdots.\end{aligned} \tag{A6}$$

This completes the proof.

[1] V. E. Zakharov and A. B. Shabat, Zh. Eksp. Teor. Fiz. **61**, 118 (1971) [Sov. Phys.-JETP **34**, 62 (1972)].
[2] S. V. Manakov, Zh. Eksp. Teor. Fiz. **65**, 1392 (1973) [Sov. Phys.-JETP **38**, 693 (1974)].
[3] H. Segur and M. J. Ablowitz, preceding paper, J. Math. Phys. **17**, 710 (1976).
[4] V. I. Karpman and V. P. Sokolov, Zh. Eksp. Teor. Fiz. **54**, 1568 (1968) [Sov. Phys.-JETP **27**, 839 (1968)].
[5] G. L. Lamb, Rev. Mod. Phys. **43**, 99 (1971).
[6] M. J. Ablowitz, D. J. Kaup, A. C. Newell and H. Segur, Stud. Appl. Math. **53**, 249 (1974).

Quantum numbers for particles in de Sitter space*

J. Patera and P. Winternitz

Centre de Recherches Mathématiques, Université de Montréal, Montréal, H3C 3J7 Canada

H. Zassenhaus†

Fairchild Distinguished Scholar, Caltech, Pasadena, California
(Received 2 June 1975)

All subalgebras of the Lie algebra of the de Sitter group O(4,1) are classified with respect to conjugacy under the group itself. The maximal continuous subgroups are shown to be O(4), O(3,1), $D \square E(3)$ (the Euclidean group extended by dilatations), and $O(2) \otimes O(2,1)$. Representatives of each conjugacy class are shown in the figures, also demonstrating all mutual inclusions. For each subalgebra we either derive all invariants (both polynomial and nonpolynomial ones) or prove that they have none. The mathematical results are used to discuss different possible sets of quantum numbers, characterizing elementary particle states in de Sitter space (or the states of any physical system, described by this de Sitter group).

I. INTRODUCTION

The aim of this paper is to study some properties of the de Sitter group O(4, 1), i.e., the group of motions of a four-dimensional space–time continuum with constant positive curvature. We shall provide a complete analysis of the continuous subgroup structure of O(4, 1), i.e., classify all subgroups into equivalence classes with respect to inner automorphisms of the group itself and construct a lattice of its continuous subgroups. We also consider the Lie algebra of each subgroup and find all its invariants, if they exist, or, as the case may be, prove that none exist. Invariants, in this article, will be defined to include Casimir operators (polynomials in the generators), harmonics (ratios of polynomials), and general nonpolynomial invariants.

The de Sitter group O(4, 1) [as well as the other de Sitter group O(3, 2)] is of considerable interest in relativistic cosmology, elementary particle theory, and also atomic physics. Indeed, the de Sitter spaces with positive or negative curvature[1] are the simplest generalizations of the flat Minkowski space–time of special relativity, capable of providing a model of the expanding universe which we live in. All laws of physics in such a universe would be invariant with respect to one of the de Sitter groups,[2,3] rather than with respect to the Poincaré group. Kinematic conservation laws (energy, linear, and angular momentum, position of the center-of-mass, etc.) will be related to the Lie algebra of the de Sitter group (and its enveloping algebra).

The de Sitter groups are of interest in elementary particle physics for several reasons. First of all, by definition, an elementary physical system in a de Sitter world would be a system described by a wavefunction transforming according to an irreducible unitary representation of the de Sitter group. Complete sets of commuting operators in the enveloping algebra of the de Sitter algebras (i.e., the Casimir operators of the entire group plus, e.g., Casimir operators of a certain chain of subgroups) will then provide the quantum numbers of such particles in definite states. A large amount of literature exists on elementary particle theory in de Sitter space, in particular dealing with problems of localization, the positivity of energy (or lack thereof), generalizations of the Dirac equation and other invariant equations, etc. (see, e.g., Refs. 4–11 and many others). A large body of work also exists on the representation theory of the de Sitter group (see, e.g., the classical papers[12–14]).

Aside from the aspect of considering particle or field theory in curved space and thus incorporating some aspects of gravitational interactions, the de Sitter world may be of interest in that it provides a possible way of avoiding the O'Raifeartaigh theorem.[15] Indeed, while it is not possible to combine the Poincaré group and an internal symmetry group, like SU(3), into a larger group, providing a discrete mass spectrum, such a unification is possible if one of the de Sitter groups is taken as the space–time group.[10,16]

From a different point of view the de Sitter group O(4, 1) is of interest in ordinary elementary particle theory in Minkowski space. Indeed, it has been shown[17] that certain canonical momentum dependent transformations of the ordinary free-particle Dirac equation exist and form an O(4, 1) group. Different subgroups of O(4, 1) then provide different specific transformations of interest, e.g., the Foldy–Wouthuysen transformation[18] is associated with an O(4) subgroup of O(4, 1).

It should also be remembered that the de Sitter groups are among the maximal subgroups of the conformal group of space–time, isomorphic to SO(4, 2). Thus, it may be of interest to consider interactions, breaking down the exact O(4, 2) symmetry, e.g., of relativistic zero-mass equations or of some conformally invariant field theory, to a de Sitter symmetry and further to lower symmetries, corresponding to subgroups of the de Sitter groups. Such reductions of conformal symmetry have been considered in the literature.[19,20]

A further reason why it is of interest to study the de Sitter groups and their subgroups is that within certain restrictions all possible "kinematical groups" can be considered to be contractions[21,22] of the de Sitter groups.[23,24] Indeed, taking the speed of light and/or the radius of curvature to infinity in various ways, we can obtain the Poincaré group, the Galilei group, and several other groups of interest. Again, knowledge of the subgroup structure of the de Sitter groups will make

it possible to systematically study contractions with respect to which certain subgroups of physical interest remain invariant.

Finally let us mention that the O(4, 1) group has made its appearance in atomic physics as one of the possible "dynamical noninvariance groups" of the hydrogen atom.[25–35] Indeed, the hydrogen atom is well known to have an O(4) symmetry group, responsible for the accidental degeneracy of its bound state levels, and an O(3, 1) symmetry for the Coulomb scattering states. Both of these can be embedded into a larger group (O(4, 1), O(4, 2), SL(4, R), etc.), the Lie algebras of which contain raising and lowering operators that do not commute with the Hamiltonian. In turn, a study of the subgroups of the corresponding invariance and noninvariance groups will provide a classification of possible symmetry breakings (e.g., by external fields).

At this point it may be appropriate to summarize the reasons why we are interested in the subgroups of the de Sitter group O(4, 1) and more generally in the subgroups of any group of interest in physics (and other applications). Indeed, we have already written four related articles. In the first[36] we found all conjugacy classes of maximal solvable subalgebras of the algebras of the pseudounitary groups SU(p, q) and all subalgebras of LSU(2, 1). In the second[37] we classified all maximal solvable subalgebras of LSO(p, q). In the third[38] and fourth[39] we provided complete lists of all classes of continuous subgroups of the Poincaré group and of the similitude group (the Poincaré group extended by dilations). The general motivation for our program was discussed previously.[36–39] In connection with the O(4, 1) group let us just stress a few points (also having general validity).

1. In a quantum theory in de Sitter space a lattice of subgroups of the de Sitter group will provide us with different complete sets of quantum numbers for elementary physical systems. It should be noted, however, that Casimir operators of continuous subgroups, while providing the simplest types of observables, by no means provide all possible sets of observables. For a discussion of nonsubgroup type observables see, e.g., Refs. 40–43.

2. A knowledge of the subgroup structure is important in group representation theory. Thus different subgroups can be used to induce representations[44] of the group and in particular provide different parametrizations of the group itself. Further, different chains of subgroups provide different bases for representations and lead to different special functions as basis functions.

3. A classification of subgroups provides a classification of different homogeneous manifolds, upon which the group acts transitively.[45] It is often of interest to construct physical wavefunctions as functions on such homogeneous manifolds (and not necessarily simply as functions on the space–time manifolds)[46–48]

4. Since most symmetries in nature are broken ones, it is of considerable interest to discuss symmetry breaking interactions, boundary conditions, etc., reducing the symmetry with respect to a group to that with respect to a subgroup.

It should be noted that we are using Lie algebraic techniques and thus can only provide a classification of continuous subgroups. From the point of view of physical applications in particular those mentioned above, discrete subgroups of Lie groups are also of very considerable interest and we plan to return to this problem. For relevant literature see, e.g., Refs. 49–51.

The subgroup structure of the Lorentz group O(3, 1) has been studied in detail (see, e.g., Refs. 40, 46), in particular in connection with two-variable expansions of relativistic scattering amplitudes. Each subgroup reduction provided a different expansion. Thus, O(3, 1) $\supset$ O(3) was related to partial wave analysis, O(3, 1) $\supset$ O(2, 1) to Regge pole theory, O(3, 1) $\supset$ E(2) to eikonal expansions. For a review of this field see Ref. 52.

In Sec. 2 of this paper we derive a list of all conjugacy classes of subalgebras of the de Sitter algebra LO(4, 1). Results are presented in figures and conjugacy is considered with respect to O(4, 1), SO(4, 1), and $SO_0(4, 1)$ [the continuous component of identity of the O(4, 1) de Sitter group] and in some cases also with respect to the subgroups themselves. In Sec. 3 we find the Casimir operators for all subgroups of O(4, 1) that have them and construct a lattice of these subgroups. We also construct coordinates on an O(4, 1) hyperboloid, allowing the separation of variables in the Laplace operator and corresponding to the individual subgroup chains. We discuss the meaning of the occurring quantum numbers. Our results and future outlook are summarized in Sec. 4.

2. CONTINUOUS SUBGROUPS OF THE DE SITTER GROUP O(4,1)

1. Definitions and general method

We shall make use of two equivalent definitions of the algebra LO(4, 1) of the group O(4, 1). Thus, the usual definition of O(4, 1) as the group of linear homogeneous transformations of a real five-dimensional space x_μ ($\mu = 0, 1, \ldots, 4$), leaving the quadratic form $x^2 = x_1^2 + x_2^2 + x_3^2 + x_4^2 - x_0^2$ invariant leads to the Lie algebras of 5×5 real matrices X satisfying

$$X^T I + IX = 0, \tag{1}$$

where

$$I = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & -1 \end{pmatrix} \tag{2}$$

and T indicates a transposed matrix. The elements g of the group O(4, 1) then satisfy

$$gIg^T = I. \tag{3}$$

This group has four components, similarly as the Lorentz group O(3, 1). Proper de Sitter transformations, constituting the group SO(4, 1), satisfy

$$\det g = 1 \tag{4}$$

in addition to (3), and proper orthochronous transformations $SO_0(4,1)$ satisfy (3), (4), and

$$g_{00} \geq 1 \tag{5}$$

[within O(4,1) we could also have $\det g = -1$ and/or $g_{00} \leq -1$].

An alternative realization of LO(4,1), also useful for our purposes, is obtain by replacing the matrix I of (2) by

$$J = \begin{pmatrix} 0&0&0&0&1 \\ 0&1&0&0&0 \\ 0&0&1&0&0 \\ 0&0&0&1&0 \\ 1&0&0&0&0 \end{pmatrix} \tag{6}$$

in Eqs. (1) and (3). In this realization the LO(4,1) matrices $\tilde{X}$ satisfy

$$\tilde{X}^T J + J\tilde{X} = 0. \tag{7}$$

We choose a basis $M_{\mu\nu}$ ($\mu, \nu = 4, 3, 2, 1, 0$) for LO(4,1) satisfying

$$[M_{\mu\nu}, M_{\sigma\tau}] = I_{\nu\sigma} M_{\mu\tau} - I_{\nu\tau} M_{\mu\sigma} + I_{\mu\tau} M_{\nu\sigma} - I_{\mu\sigma} M_{\nu\tau} \tag{8}$$

with $I_{ik} = \delta_{ik}$, $i, k = 4, 3, 2, 1$, $I_{00} = -1$, and $I_{0i} = I_{i0} = 0$. In the realization (1) this basis consists of the matrices

$$M_{ik} = Y_{ik} - Y_{ki} \quad \text{and} \quad M_{i0} = M_{0i} = Y_{0i} + Y_{i0}, \quad i, k = 4, 3, 2, 1. \tag{9}$$

The matrices $Y_{\mu\nu}$ have 1 on the intersection of the μth row and νth column and zeros elsewhere.

The matrices $\tilde{X}$ of realization (7) are related to those of realization (1) by the transformation

$$\tilde{X} = ZXZ^{-1}, \tag{10}$$

where

$$Z = \begin{pmatrix} 1/\sqrt{2} & 0&0&0 & -1/\sqrt{2} \\ 0&1&0&0&0 \\ 0&0&1&0&0 \\ 0&0&0&1&0 \\ 1/\sqrt{2} & 0&0&0 & 1/\sqrt{2} \end{pmatrix}, \quad Z^{-1} = Z^T. \tag{11}$$

Making use of either of these realizations, we distinguish two types of subalgebras of LO(4,1), namely those imbedded irreducibly and those imbedded reducibly in LO(4,1). The irreducibly imbedded ones, by definition, do not leave any nontrivial real subspace in the O(4,1) space invariant. It can be shown[53] that the LO(4,1) algebra [contrary to the LO(3,2) algebra] has no subalgebras of this type. Thus, we only have to classify all reducible subalgebras, and we start out by finding all maximal subalgebras of LO(4,1). To do this, we simply consider a representative of each conjugacy class of subspaces of the O(4,1) space [conjugacy is considered under O(4,1)] and find the subalgebra that leaves this space invariant. We then find all subalgebras of each maximum subalgebra and we can make use of methods and results obtained earlier.[36–39,54]

II. Maximal subgroups of the O(4,1) de Sitter group

A general LO(4,1) matrix in realization (1) can be written as

$$X = \begin{pmatrix} 0 & a & b & c & d \\ -a & 0 & e & f & g \\ -b & -e & 0 & h & j \\ -c & -f & -h & 0 & k \\ d & g & j & k & 0 \end{pmatrix}. \tag{12}$$

In realization (7) we have

$$\tilde{X} = ZXZ^{-1} = \begin{pmatrix} -d & \frac{a-g}{\sqrt{2}} & \frac{b-j}{\sqrt{2}} & \frac{c-k}{\sqrt{2}} & 0 \\ -\frac{a+g}{\sqrt{2}} & 0 & e & f & -\frac{a-g}{\sqrt{2}} \\ -\frac{b+j}{\sqrt{2}} & -e & 0 & h & -\frac{b-j}{\sqrt{2}} \\ -\frac{c+k}{\sqrt{2}} & -f & -h & 0 & -\frac{c-k}{\sqrt{2}} \\ 0 & \frac{a+g}{\sqrt{2}} & \frac{b+j}{\sqrt{2}} & \frac{c+k}{\sqrt{2}} & d \end{pmatrix}. \tag{13}$$

Let us now consider subspaces of the five-dimensional space of real vectors $(x_4, x_3, x_2, x_1, x_0)$ with metric $x_4^2 + x_3^2 + x_2^2 + x_1^2 - x_0^2 = \text{inv}$. The subspaces will differ by their dimension and signature.

A. One-dimensional subspaces

A1. *Timelike subspace* [*signature* (−)]: Consider the space T generated by the column vector (00001) (which we write in row form to save space) and require that the operator X of (12) leaves if invariant: $XT \subseteq T$. This implies $d = g = j = k = 0$ and we obtain the algebra LO(4) of the four-dimensional rotation group, generated by

$$M_{ik} \text{ with } i, k = 4, 2, 3, 1. \tag{14}$$

A2. *Spacelike subspace* [*signature* (+)]: Consider the space S generated by the vector (10000) (which again should be a column) and require that it be invariant under (12). This implies that $a = b = c = d = 0$ and we obtain the algebra LO(3,1) of the homogeneous Lorentz group, generated by

$$L_1 = M_{32}, \quad L_2 = -M_{31}, \quad L_3 = M_{21}, \quad K_1 = M_{10}, \quad K_2 = M_{20}, \quad \text{and} \quad K_3 = M_{30}. \tag{15}$$

A3. *Lightlike subspace* [*signature* (0)]: Consider the space L, generated by the vector (1000-1) in the realization (1). Applying operator Z to it, we obtain

$$\tilde{S} = ZS \simeq \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

in realization (7). Requiring that $\tilde{S}$ be invariant under $\tilde{X}$ of (13), we find $g = -a$, $j = -b$, $k = -c$, i.e., we obtain a seven-parameter algebra, generated by

$$D \equiv M_{40}, \quad L_1 \equiv M_{32}, \quad L_2 \equiv -M_{31}, \quad L_3 \equiv M_{21}, \quad P_1 \equiv M_{41} - M_{10}, \quad P_2 \equiv M_{42} - M_{20}, \quad \text{and} \quad P_3 \equiv M_{43} - M_{30}. \tag{16}$$

These generators satisfy the commutation relations

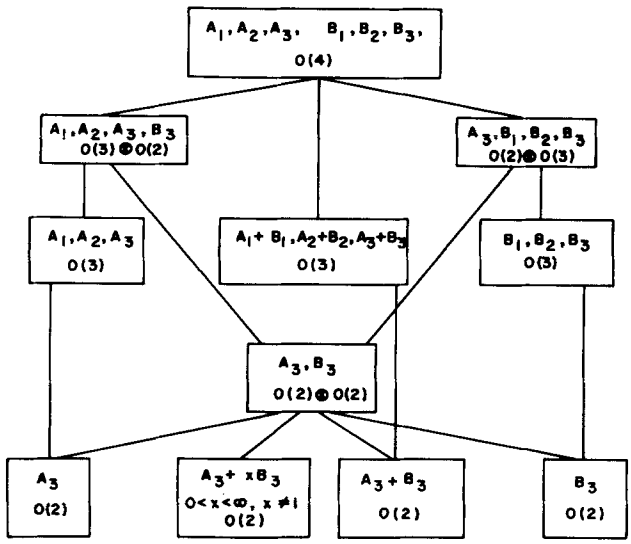

FIG. 1. SO(4) conjugacy classes of subalgebras of LO(4) and the groups they generate.

$$[L_i, L_k] = \epsilon_{ikl} L_l, \quad [L_i, P_k] = \epsilon_{ikl} P_l,$$
$$[D, L_i] = 0, \qquad [D, P_i] = -P_i. \tag{17}$$

Thus, we obtain the algebra D□LE(3), i.e., the algebra of the Euclidean group in three-dimensions, extended by dilations.

B. Two-dimensional subspaces

*B*1. *Signature* (++): We use realization (1), consider the space $(SS) = (x_4, x_3, 0, 0, 0)$, and require that $X(SS) \subseteq (SS)$. This implies $b = c = d = e = f = g = 0$ in (12), i.e., we obtain the algebra LO(2) ⊕ LO(2, 1), generated by

$$L_3 \equiv M_{21}, \quad K_1 \equiv M_{10}, \quad K_2 \equiv M_{20}, \quad \text{and} \quad A \equiv M_{43} \tag{18}$$

with

$$[L_3, K_1] = K_2, \quad [K_2, L_3] = K_1, \quad [K_1, K_2] = -L_3,$$
$$[A, K_1] = [A, K_2] = [A, L_3] = 0. \tag{19}$$

*B*2. *Signature* (+−): Using realization (1), we require that the space $(ST) = (x_4, 0, 0, 0, x_0)$ be invariant. This implies that $a = b = c = g = j = k = 0$ in (12), i.e., we obtain the algebra LO(3) ⊕ LO(1, 1), generated by

$$L_1 = M_{32}, \quad L_2 = -M_3, \quad L_3 = M_{21}, \quad \text{and} \quad D = M_{40}. \tag{20}$$

However, any transformation leaving space ST invariant also leaves space L invariant and indeed we see that the subalgebra (20) is contained in (17) and is hence not maximal.

*B*3. *Signature* (+0): We use realization (1) of LO(4, 1) and require that the space $(x, y, 0, 0, -x)$ remains invariant. This implies $e = f = 0$, $g = -a$, $j = -b$, $k = -c$. This leads us to the subalgebra $(D \oplus LO(2)) \square T_3$, generated by

$$L_3 = M_{21}, \quad D = M_{40}, \quad P_1 = M_{41} - M_{10},$$
$$P_2 = M_{42} - M_{20}, \quad P_3 = M_{43} - M_{30}. \tag{21}$$

Again, this subalgebra is contained in (17) and is hence not maximal.

We have thus considered all one- and two-dimensional subspaces of the O(4, 1) space that are not conjugate under O(4, 1). Subalgebras that leave three- or four-dimensional subspaces invariant will automatically also leave their two- or one-dimensional orthogonal complements invariant and will hence coincide with those obtained above, or be contained in them.

To summarize, the algebra LO(4, 1) has exactly four O(4, 1) conjugacy classes of maximal subalgebras, namely LO(4) of (14), LO(3, 1) of (15), D□LE(3) of (16), and LO(2) ⊕ LO(2, 1) of (18). Since an arbitrary one- or two-dimensional subspace of the O(4, 1) space with the corresponding signature can be transformed into one of the spaces S, T, L or (SS) by an $SO_0(4, 1)$ transformation the above algebras also represent all $SO_0(4, 1)$ classes of maximal subalgebras.

We now proceed to classify all subalgebras of each maximal subalgebra.

III. Subalgebras of LO(4)

The algebra LO(4) is isomorphic to LO(3) ⊕ LO(3). Its subalgebras can be obtained using the "Goursat twist method" and were originally classified by Goursat.[55] The method, in application to Lie algebras, was discussed in a previous publication,[38] so here we omit all details.

Let us introduce the notation

$$A_1 = \tfrac{1}{2}(M_{32} + M_{41}), \quad A_2 = \tfrac{1}{2}(-M_{31} + M_{42}),$$
$$A_3 = \tfrac{1}{2}(M_{21} + M_{43}),$$
$$B_1 = \tfrac{1}{2}(M_{32} - M_{41}), \quad B_2 = \tfrac{1}{2}(M_{31} + M_{42}),$$
$$B_3 = \tfrac{1}{2}(-M_{21} + M_{43}), \tag{22}$$

so that

$$[A_i, A_k] = \epsilon_{ikl} A_l, \quad [B_i, B_k] = \epsilon_{ikl} B_l, \quad [A_i, B_k] = 0. \tag{23}$$

The algebra LO(4) ~ LO(3) ⊕ LO(3) will have two types of subalgebras. The first type, "nontwisted subalgebras," are obtained by taking direct sums of subalgebras of the one LO(3) with those of the other. The second type, "twisted subalgebras," involve generators that are not conjugate to either A_i nor B_i. Only two such subalgebras exist: the LO(3) algebra generated by $A_1 + B_1$, $A_2 + B_2$, and $A_3 + B_3$, contained reducibly in LO(4) and a one-dimensional subalgebra $A_3 + xB_3$, depending on one parameter x. An SO(4) transformation changing $A_3 + xB_3$ into $A_3 - xB_3$ can easily be constructed; hence we can take $0 < x < \infty$.

A lattice of SO(4) conjugacy classes of subalgebra of LO(4) is given in Fig. 1.

If conjugacy is considered under O(4), rather than SO(4), then A_i and B_i are conjugate

$$gA_i g^{-1} = B_i,$$

$$g = \begin{pmatrix} 1 & & & \\ & -1 & & \\ & & -1 & \\ & & & -1 \end{pmatrix}. \tag{24}$$

The lattice of Fig. 1 simplifies under O(4), in that all

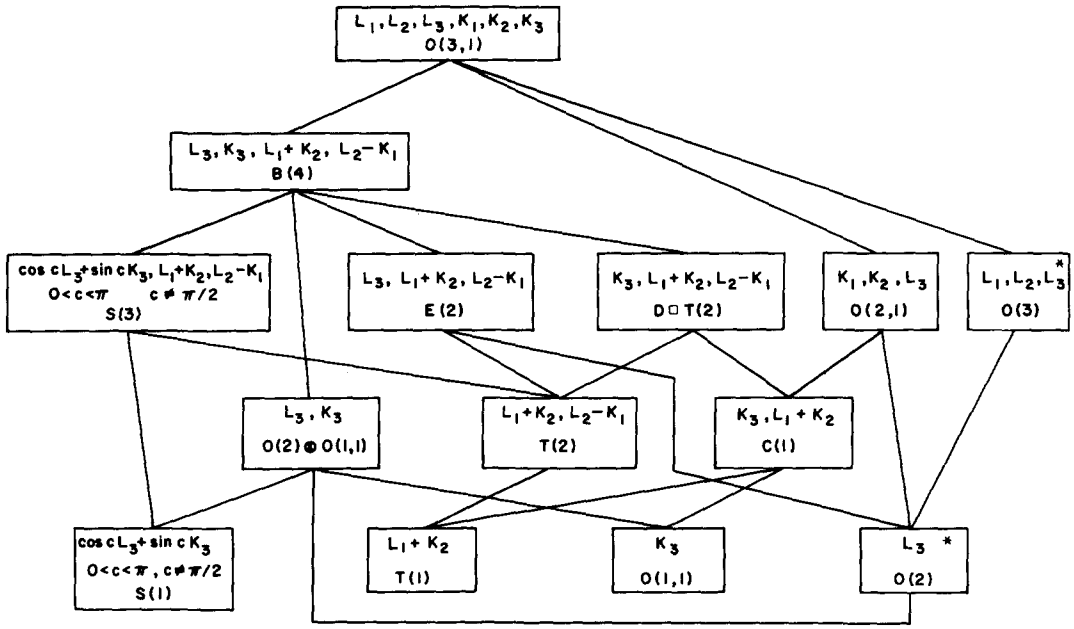

FIG. 2. $SO_0(3,1)$ conjugacy classes of subalgebras of LO(3, 1). The group generated by the algebra is also given. Here B(4) indicates the Borel subgroup, i.e., the maximal solvable subgroup of O(3, 1); the S in S(3) and S(1) stands for "screw" (a combination of a rotation about an axis with a boost along the same axis; T(1) stands for translations along one axis, E(2) for the Euclidean group, and D for dilations. An asterisk indicates subalgebras conjugate to ones contained in Fig. 1.

algebras in the furthest to right column become conjugate to those in the left-hand column and the parameter x can be restricted to $0<x<1$.

Imbedding O(4) into O(4, 1), we replace (24) by

$$g=\begin{pmatrix} 1 & & & & \\ & -1 & & & \\ & & -1 & & \\ & & & -1 & \\ & & & & -1 \end{pmatrix}. \tag{25}$$

with $\det g=1$, $g_{00}=-1$. Thus g is contained in SO(4, 1), but not in $SO_0(4,1)$.

IV. Subalgebras of LO(3,1)

The subalgebras of LO(3, 1) are known.[40] For completeness we give a lattice of $SO_0(3,1)$ classes of subalgebras of LO(3, 1) in Fig. 2. If we consider conjugacy under O(3, 1), i.e., include parity, then $0<c<\pi/2$ in S(3) and S(1). This transformation is contained in $SO_0(4,1)$. Note that the algebras LO(3) and LO(2) in Fig. 2 are already contained in Fig. 1.

V. Subalgebras of D □ LE(3)

The continuous subgroups of the Euclidean group E(3) have been classified earlier.[54] The E(3) conjugacy classes of subalgebras of LE(3) are given in Fig. 3.

If we add parity to E(3) then $a>0$ in L_3+aP_3 and $\{L_3+aP_3, P_1, P_2\}$. The corresponding transformation

$$g=\begin{pmatrix} -1 & & & & \\ & 1 & & & \\ & & 1 & & \\ & & & 1 & \\ & & & & -1 \end{pmatrix}. \tag{26}$$

is contained in SO(4, 1) but not in $SO_0(4,1)$.

Let us now add the dilations, generated by D, to E(3). Since $[D, L_3]=0$ but $[D, P_3]=-P_3$, we can transform the parameter a in L_3+aP_3 into $a=1$. All subalgebras of Fig. 3 (with $a=1$) will then also be subalgebras of $D\,\square\,\mathrm{LE}(3)$ and none of them will be conjugate to each other. Further subalgebras will involve D and will be of two types. The first type is obtained by simply adding D as a generator to all subalgebras of E(3) that split over their intersection with the translations (i.e., subalgebras not containing the generator L_3+P_3). The second type of algebra can be written as

$$\{D+a_iL_i+x_iP_i; E_\alpha\}, \tag{27}$$

where E_α is one of the subalgebras of LE(3). We must now run through all subalgebras E_α, spell out the additional generator $\tilde{D}=D+a_iL_i+x_iP_i$, and set all a_i and x_i equal to zero, if the corresponding L_i and P_i are contained in E_α. Then we simplify $\tilde{D}$ further, using trans-

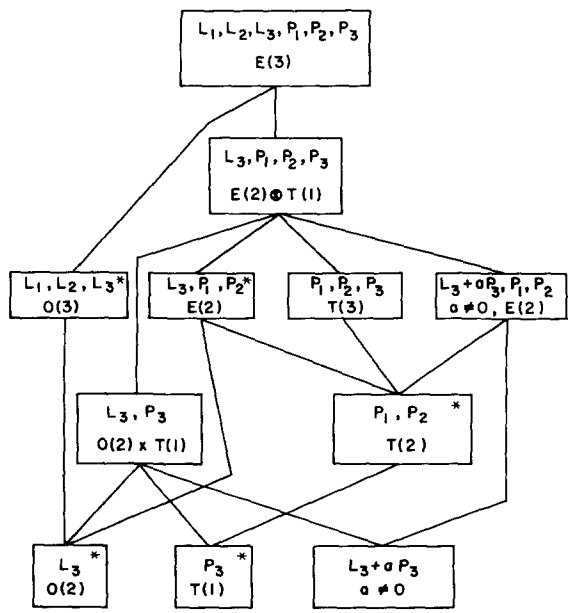

FIG. 3. E(3) conjugacy classes of subalgebras of LE(3) and the groups they generate. An asterisk indicates subalgebras conjugate to ones contained in Figs. 1 or 2.

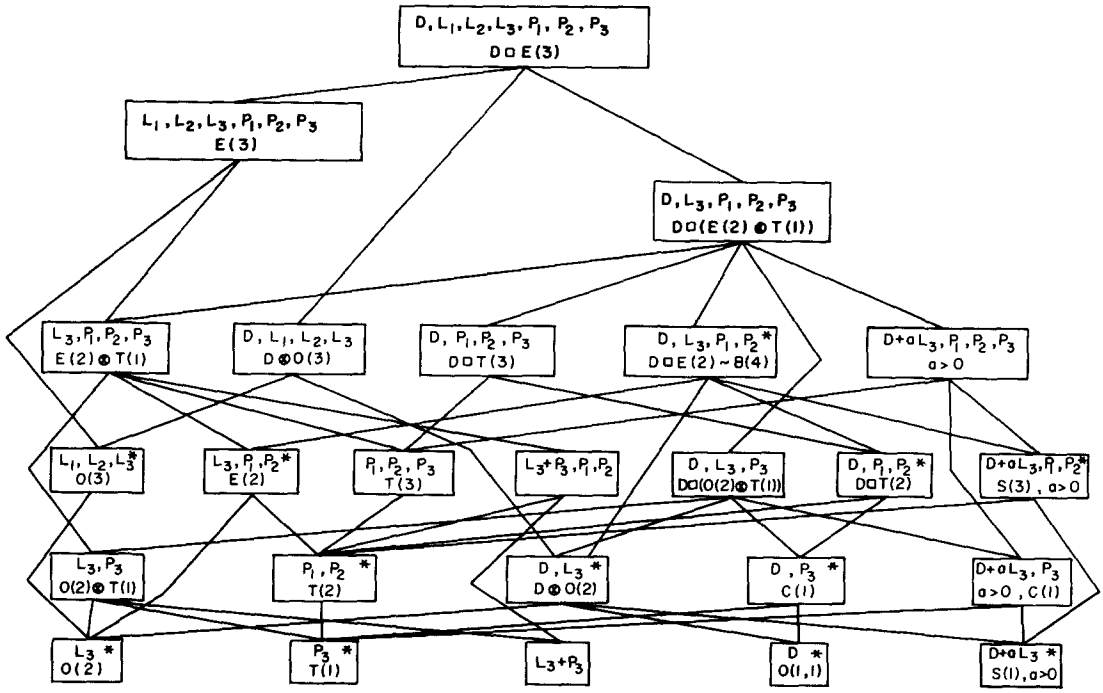

FIG. 4. Subalgebras of $D \square LE(3)$ and the groups they generate. An asterisk indicates subalgebras conjugate to ones contained in Figs. 1 or 2.

formations leaving E_α invariant and finally we enforce that D and E_α together should form a closed algebra. The last two steps can be performed in an elegant manner, using cohomology theory.[38,39] In the present case the task is so simple that we just proceed in a straightforward manner. Note that new subalgebras are obtained only if at least one of the a_i or x_i is nonzero.

The only subalgebras of LE(3) leading to such nontrivial subalgebras of $D \square LE(3)$ are those not containing any rotations, i.e., $\{P_1, P_2, P_3\}$, $\{P_1, P_2\}$, $\{P_3\}$, and $\{0\}$. Consider them individually.

a. $\{P_1, P_2, P_3\}$. Write the additional element as $\tilde{\tilde{D}} = D + a_i L_i$. Performing an O(3) rotation, we can transform $\tilde{\tilde{D}}$ into $\tilde{D} = D + aL_3$, $a > 0$. We obtain the algebra

$$D + aL_3, \quad P_1, P_2, P_3, \quad a > 0. \tag{28}$$

b. $\{P_1, P_2\}$. We write $\tilde{D} = D + a_i L_i + xP_3$. Since $[D, P_3] = -P_3$, the group transformation $\exp(yP_3)$, contained in E(3), can be used to transform x into zero. The commutation relations $[\tilde{D}, P_1] = -P_1 - a_2 P_3 + a_3 P_2$ and $[\tilde{D}, P_2] = -P_2 + a_1 P_3 - a_3 P_1$ imply $a_1 = a_2 = 0$ and a rotation through π about axis 1 or 2 can be used to change the sign of a_3. We find the algebra

$$D + aL_3, \quad P_1, P_2, \quad a > 0. \tag{29}$$

c. $\{P_3\}$. We write $\tilde{D} = D + a_i L_i + x_1 P_1 + x_2 P_2$. The transformation $\exp(xP_1)$ and $\exp(yP_2)$ can be used to transform x_1 and x_2 into zero. The relation $[\tilde{D}, P_3] = -P_3 - a_1 P_2 + a_2 P_1$ implies $a_1 = a_2 = 0$. A rotation through π about axis 1 or 2 will change the sign of a_3. We find

$$D + aL_3, \quad P_3, \quad a > 0. \tag{30}$$

d. $\{0\}$. We have $\tilde{D} = D + a_i L_i + x_i P_i$. A rotation can be used to obtain $a_1 = a_2 = 0$. Transformations of the type $\exp(aL_3)$ and $\exp(y_i P_i)$ can be used to obtain $x_1 = x_2 = x_3 = 0$. We obtain the algebra

$$D + aL_3, \quad a > 0. \tag{31}$$

None of the obtained subalgebras of $D \square E(3)$ can be further simplified by O(4, 1) transformations. The conjugacy classes [under the Euclidean group extended by dilatations and parity and also under O(4, 1)] are summarized in Fig. 4.

All algebras of Fig. 4 leave a one-dimensional lightlike space L invariant. Many of them also leave a one-dimensional timelike or spacelike vector space invariant and are thus contained in LO(4) or LO(3, 1). This can easily be established for each subalgebra separately. We denote by an asterisk in Fig. 4 those subalgebras that are conjugate, under O(4, 1), to algebras in Figs. 1 or 2.

VI. Subalgebras of LO(2) ⊕ LO(2,1)

All subalgebras of this algebra can be obtained either as the direct sums of subalgebras of LO(2) and LO(2, 1) (including the trivial ones) or by applying the Goursat twist method. The results are given in Fig. 5. All of the subalgebras in Fig. 5 also leave a one-dimensional vector space invariant and will thus already have been listed in Figs. 1, 2, or 4. This can easily be established by inspection, and we indicate all previously listed subalgebras by an asterisk in Fig. 5.

VII. All subalgebras of LO(4,1)

Figures 1, 2, 4, 5 can now be used to compile a complete lattice of subalgebras of LO(4, 1) presented in Fig. 6. We consider conjugation under O(4, 1), so as to keep the size of Fig. 6 manageable.

This completes our investigation of the subalgebras of LO(4, 1) and thus also of the continuous subgroups of O(4, 1).

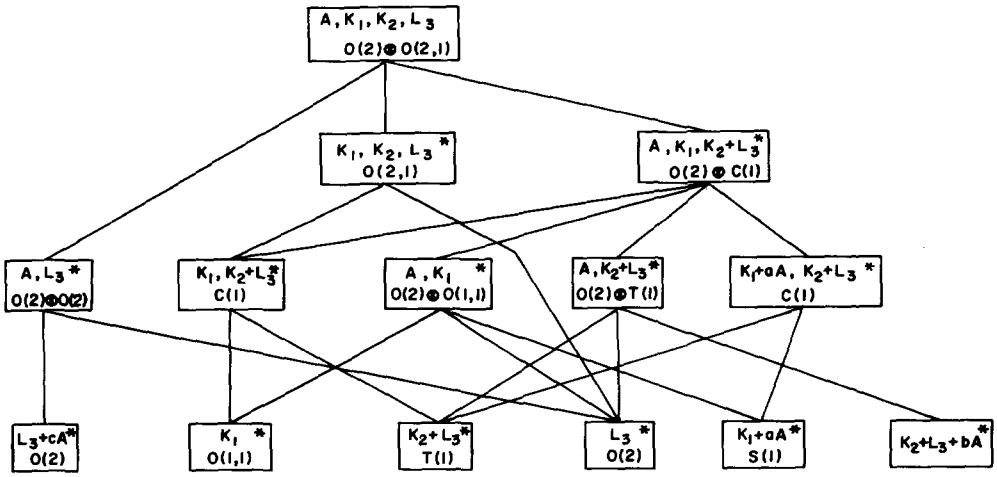

FIG. 5. Subalgebras of LO(2) $\oplus$LO(2, 1). If conjugacy is considered under S[O(2) $\times O_0(2,1)$] we have $a \neq 0$, $b \neq 0$, $c \neq 0$. Conjugacy under S[O(2) $\times O(2,1)$] gives $a > 0$, $b > 0$, $c > 0$. Conjugacy under $SO_0(4,1)$ gives $a \neq 0$, $b = \pm 1$, $c \neq 0$. Conjugacy under SO(4, 1) gives $a > 0$, $b = 1$, $c > 0$. An asterisk indicates algebras conjugate under O(4, 1) to algebras in Figs. 1, 2, or 4.

3. INVARIANTS OF SUBALGEBRAS OF LO(4,1)

I. General method for finding invariants of Lie algebras

Having thus provided a classification of all subalgebras of LO(4, 1), we now wish to determine which of the subalgebras have invariants, in particular Casimir operators and to find all of them. These will then provide us with observables and quantum numbers for particles in a de Sitter space [or for any physical system for which O(4, 1) is a relevant symmetry group].

Let us briefly discuss our method of searching for invariants. Consider the Lie algebra $\mathcal{L}$ generated by the operators $A_1, \ldots, A_n$, satisfying

$$[A_i, A_k] = \sum_{l=1}^{n} f_{ik}^{l} A_l. \tag{32}$$

We shall represent the generators A_i as differential operators acting on functions $F(a_1, \ldots, a_n)$ of n variables (n is the dimension of the algebra). Indeed, if we put

$$A_i = \sum_{k,l=1}^{n} f_{ik}^{l} a_l \frac{\partial}{\partial a_k}, \quad i = 1, \ldots, n, \tag{33}$$

the operators A_i will satisfy (32). We are now interested in finding an operator valued function $P(A_1, \ldots, A_n)$, commuting with all A_i. This is equivalent to finding a numerical function $P(a_1, \ldots, a_n)$, annihilated by all generators (33):

$$A_i P(a_1, \ldots, a_n) = 0, \quad i = 1, \ldots, n, \tag{34}$$

then symmetrizing P with respect to all permutations of a_i and replacing the variables a_i by the operators A_i.

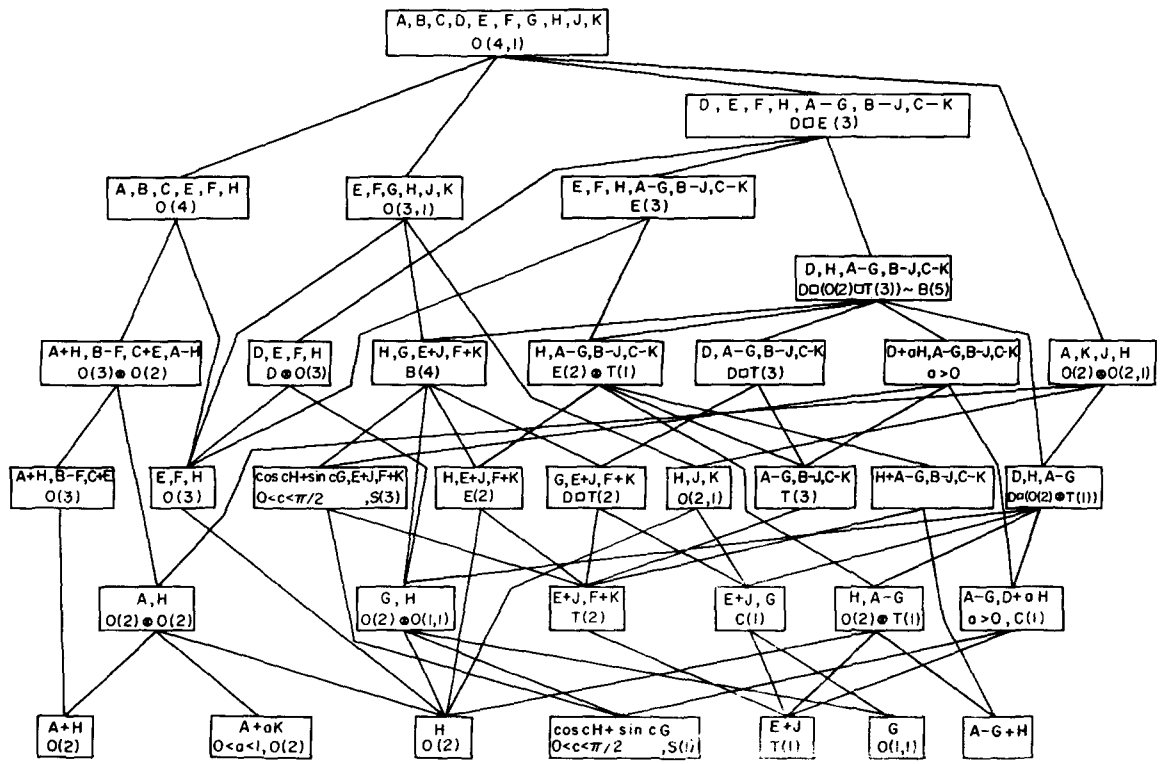

FIG. 6. Subalgebras of LO(4, 1) and the groups they generate. We use the notation $A = M_{43}$, $B = M_{42}$, $C = M_{41}$, $D = M_{40}$, $E = M_{32}$, $F = M_{31}$, $G = M_{30}$, $H = M_{21}$, $J = M_{20}$, and $K = M_{10}$. Conjugacy is considered under O(4, 1).

We thus reduce the search for Casimir operators to the problem of solving the system of homogeneous linear partial differential equations (34). If the system is contradictory, i.e., does not have a nonzero solution, then the algebra has no Casimir operators. On the other hand, solutions may exist, but not be expressible in terms of polynomials in a_i. We then obtain "generalized Casimir operators," i.e., operators not lying in the enveloping algebra of L but still commuting with all generators and hence having a fixed numerical value within each irreducible representation of the algebra. If polynomial solutions of (34) exist, they will provide us with Casimir operators. Generally speaking, among such solutions we must find an integrity basis, i.e., a minimal set of operators C_i such that any invariant can be expressed as a polynomial in C_i.

Note that for semisimple algebras the problem is solved—the number of Casimir operators is equal to the rank of the algebra and they are all known.

II. Invariants of the subalgebras of LO(4,1)

Let us go through the algebras of Figs. 1–6 and find their Casimir operators. The Casimir operators of O(4, 1) itself are well known, namely

$$C^{(2)} = M_{\alpha\beta} I_{\beta\gamma} M_{\gamma\delta} I_{\delta\alpha},$$
$$C^{(4)} = M_{\alpha\beta} I_{\beta\gamma} M_{\gamma\delta} I_{\delta\epsilon} M_{\epsilon\xi} I_{\xi\eta} M_{\eta\xi} I_{\xi\alpha}.$$

A. LO(4) and its subalgebras

All algebras in Fig. 1 have invariants and they are quite obvious. Thus LO(4) itself has two Casimir operators $\mathbf{A}^2 = A_1^2 + A_2^2 + A_3^2$ and $\mathbf{B}^2 = B_1^2 + B_2^2 + B_3^2$. For a one-dimensional subalgebra the generator itself is an invariant; for an Abelian subalgebra all generators are invariants. The "twisted" LO(3) $(A_1 + B_1,\ A_2 + B_2,\ A_3 + B_3)$ has the Casmimir operator $(\mathbf{A} + \mathbf{B})^2$. The invariants of a direct sum of algebras will be the invariants of each component.

B. LO(3,1) and its subalgebras

The invariants of the subalgebras of LO(3, 1) are known.[40] Thus, using the notations of Fig. 2, we have the following. LO(3, 1) itself has two Casimir operators $\mathbf{L}^2 - \mathbf{K}^2$ and $\mathbf{L} \cdot \mathbf{K}$. The algebras LE(2), LO(2, 1) and LO(3) have the invariants $(L_1 + K_2)^2 + (L_2 - K_1)^2$, $K_1^2 + K_2^2 - L_3^2$, and $L_1^2 + L_2^2 + L_3^2$, respectively. The generators of Abelian or one-dimensional algebras are themselves invariants. The algebras B, S(3), $D \square (\mathrm{LT}(1) \oplus \mathrm{LT}(1))$, and C(1) have no Casimir operators. However, using the method discussed above, we can show that both S(3) and $D \square (\mathrm{LT}(1) \oplus \mathrm{LT}(1))$ have a nonpolynomial invariant. Indeed, consider the algebra $D = K_3,\ P = L_1 + K_2,\ Q = -L_2 + K_1$. We have

$$[D, P] = P, \quad [D, Q] = Q, \quad [P, Q] = 0$$

so that

$$D = p\frac{\partial}{\partial p} + q\frac{\partial}{\partial q}, \quad P = -p\frac{\partial}{\partial d}, \quad Q = -q\frac{\partial}{\partial d}. \tag{35}$$

Consider the function $F(p, q, d)$ and require

$$DF = PF = QF = 0.$$

The last two equations imply that F does not depend on d, the first implies that F is an arbitrary function of p/q. Hence, the invariant is the operator

$$X = (L_1 + K_2)/(-L_2 + K_1), \tag{36}$$

which generally speaking is not a well-defined operator. Similarly, consider S(3), generated by

$$R = \cos\phi L_3 + \sin\phi K_3, \quad P = L_1 + K_2, \quad Q = -L_2 + K_1,$$
$$0 < \phi < \pi/2 \quad \text{or} \quad \pi/2 < \phi < \pi.$$

Using the commutation relations for S(3) we write

$$R = (q\cos\phi + p\sin\phi)\frac{\partial}{\partial p} + (q\sin\phi - p\cos\phi)\frac{\partial}{\partial p},$$
$$P = -(q\cos\phi + p\sin\phi)\frac{\partial}{\partial r}, \quad Q = -(q\sin\phi - p\cos\phi)\frac{\partial}{\partial r}. \tag{37}$$

Requiring $RF = PF = QF = 0$, we obtain a nonpolynomial invariant

$$I = (P^2 + Q^2)\left(\frac{P - iQ}{P + iQ}\right)^{i\tan\phi}$$
$$= (P^2 + Q^2)\exp[2\tan\phi \cdot \arctan(Q/P)]. \tag{38}$$

C. D □ LE(3) and its subalgebras

Consider first the algebra $D \square \mathrm{LE}(3)$ itself. The invariant $F(p_1, p_2, p_3, l_1, l_2, l_3, d)$ could depend on seven variables; however, rotational invariance $L_i F = 0$ implies that it only depends on O(3) scalars $\mathbf{p}^2$, $\mathbf{l}^2$, $\mathbf{p} \cdot \mathbf{l}$, and d. Scale invariance $DF = 0$ implies that F only depends on $\mathbf{l}^2$ and $\mathbf{p}^2/(\mathbf{p} \cdot \mathbf{l})^2$. Finally translational invariance $P_i F = 0$ implies that F depends only on $\mathbf{p}^2/(\mathbf{p} \cdot \mathbf{l})^2$. Thus $D \square \mathrm{LE}(3)$ has no Casimir operators, but has one nonpolynomial invariant, the harmonic

$$I = \mathbf{P}^2/(\mathbf{P} \cdot \mathbf{L})^2. \tag{39}$$

The algebra LE(3) has the two well-known Casimir operators $\mathbf{P}^2$ and $(\mathbf{P} \cdot \mathbf{L})$, i.e., the energy and helicity of a nonrelativistic particle.

The algebra $\{D, L_3, P_1, P_2, P_3\}$ can be shown to have only one invariant, namely $(P_1^2 + P_2^2)/P_3^2$, which again is nonpolynomial.

The algebra $\{L_3, P_1, P_2, P_3\}$ has two Casimir operators: $P_1^2 + P_2^2$ and P_3.

The algebra $\{D, L_1, L_2, L_3\}$ has two Casimir operators: D and $\mathbf{L}^2$.

The algebra $\{D, P_1, P_2, P_3\}$ has no Casimir operators but two nonpolynomial invariants P_1/P_3 and P_2/P_3.

The algebra $\{D + aL_3, P_1, P_2, P_3\}$ is somewhat more complicated. We put

$$R = D + aL_3 = (-p_1 + ap_2)\frac{\partial}{\partial p_1} - (p_2 + ap_1)\frac{\partial}{\partial p_2} - p_3\frac{\partial}{\partial p_3},$$
$$P_1 = (p_1 - ap_2)\frac{\partial}{\partial r}, \quad P_2 = (p_2 + ap_1)\frac{\partial}{\partial r}, \quad P_3 = p_3\frac{\partial}{\partial r}. \tag{40}$$

Requiring $P_i F = 0$ implies that $F = F(p_1, p_2, p_3)$. Requiring $RF = 0$ implies

$$\frac{dp_1}{p_1 - ap_2} = \frac{dp_2}{p_2 + ap_1} = \frac{dp_3}{p_3}. \tag{41}$$

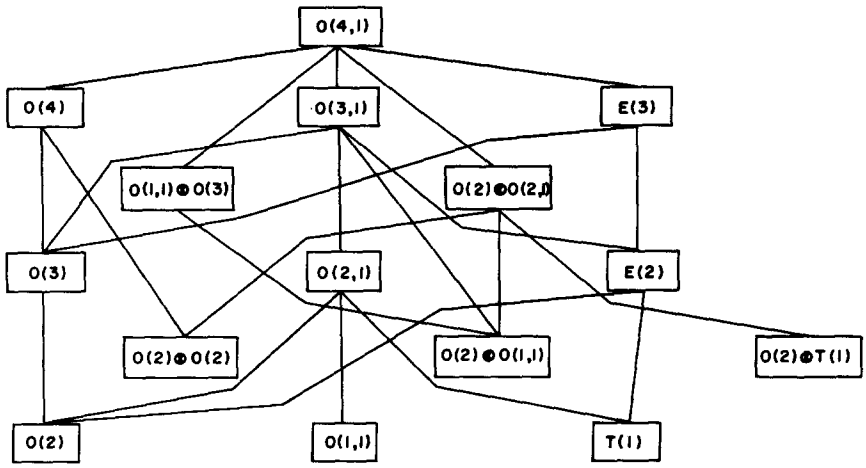

FIG. 7. Subgroups of O(4, 1) that have Casimir operators.

These equations are solved by standard methods,[56] and we obtain two (real) nonpolynomial invariants

$$X_1 = P_3\{(P_1 + iP_2)^{-(1-ia)/(1+a^2)} + (P_1 - iP_2)^{-(1+ia)/(1+a^2)}\},$$
$$X_2 = iP_3\{(P_1 + iP_2)^{-(1-ia)/(1+a^2)} - (P_1 - iP_2)^{-(1+ia)/(1+a^2)}\}. \tag{42}$$

The invariants of all other subalgebras are obvious (or have been obtained above), as are those of subalgebras of $LO(2) \oplus LO(2, 1)$.

The situation is best summarized by the diagram of Fig. 7, where we give a lattice of subgroups of O(4, 1), listing only those which have Casimir operators. More specifically, we only list a subgroup if it has a new Casimir operator, that is not also a Casimir operator for a larger subgroup, higher in the chain.

III. Quantum numbers

All chains of subgroups providing us with a complete set of observables are shown on Figs. 6 and 7, and we see 16 possible independent sets. In addition to the two Casimir operators of LO(4, 1), characterizing the system as such, we have the following possible choices of operators, characterizing the particle states.

A. Reduction to O(4)

Using the notations (22), we see that the complete set of commuting operators would contain:

$$\mathbf{A}^2, \quad \mathbf{B}^2, \tag{43}$$

and either

$$(\mathbf{A} + \mathbf{B})^2 \text{ and } A_3 + B_3 \tag{44a}$$

or

$$A_3 \text{ and } B_3. \tag{44b}$$

B. Reduction to O(3,1)

Using the notations (15) we write the observables

$$\mathbf{L}^2 - \mathbf{K}^2 \text{ and } \mathbf{L} \cdot \mathbf{K}, \tag{45}$$

supplemented by one of the following pairs:

$$\mathbf{L}^2 \text{ and } L_3, \tag{46a}$$
$$K_1^2 + K_2^2 - L_3^2 \text{ and } L_3, \tag{46b}$$
$$K_1^2 + K_2^2 - L_3^2 \text{ and } K_1, \tag{46c}$$
$$K_1^2 + K_2^2 - L_3^2 \text{ and } K_2 + L_3, \tag{46d}$$
$$(L_1 + K_2)^2 + (L_2 - K_1)^2 \text{ and } L_3, \tag{46e}$$
$$(L_1 + K_2)^2 + (L_2 - K_1)^2 \text{ and } L_1 + K_2, \tag{46f}$$
$$L_3 \text{ and } K_3. \tag{46g}$$

C. Reduction to E(3)

Using the notations (16), we write the observables as

$$\mathbf{P}^2 \text{ and } \mathbf{P} \cdot \mathbf{L}, \tag{47}$$

supplemented by one of the following pairs:

$$\mathbf{L}^2 \text{ and } L_3, \tag{48a}$$
$$P_1^2 + P_2^2 \text{ and } L_3, \tag{48b}$$
$$P_1^2 + P_2^2 \text{ and } P_3 \text{ (or } P_1, P_2 \text{ and } P_3). \tag{48c}$$

D. Reduction to O(3) × O(1,1)

This reduction, using notations (20), provides us with three quantum numbers:

$$\mathbf{L}^2, \quad L_3, \text{ and } D. \tag{49}$$

E. Reduction to O(2) ⊕ O(2,1)

Using notations (18), we obtain three quantum numbers given by

$$A, \quad K_1^2 + K_2^2 - L_3^2 \tag{50a}$$

and one of the three operators

$$L_3, \quad K_1, \text{ or } K_2 + L_3. \tag{50b}$$

The physical interpretation of each set of observables is open to discussion and depends on the specific physical system considered. We have however clarified the group theoretical significance of each set.

Several comments are in order: (i) While the "canonical" reductions of O(4, 1) to O(4), O(3, 1), and E(3) provide us with complete sets of observables, the reductions to $O(3) \otimes O(2)$ and $O(2) \otimes O(2, 1)$ provide only three quantum numbers and we are faced with a "missing

label problem." These have been discussed extensively in the literature in other connections, in particular in relation to the $SU(3) \supset O(3)$ reduction. One way to provide the missing quantum number, completely specifying the states, would be to add a further operator to the set (49) or (50). This would have to lie in the enveloping algebra of $LO(4,1)$, not, however, of the subgroup $O(3) \otimes O(1,1)$ [or $O(2) \otimes O(2,1)$] and be a scalar with respect to the corresponding subgroup. For a discussion of this problem, see Refs. 57, 58 and references therein.

(ii) We have not touched upon "nonsubgroup" type observables, i.e., complete sets of operators that can specify a state, but are not Casimir operators of any Lie subgroup. Examples of such observables in connection with O(3) and other little groups of the Poincaré group have been studied.[40–42] They can be related to discrete subgroups of the corresponding group—a question that is itself of considerable interest.

(iii) A question that to our knowledge has received no attention at all is the significance of invariants of Lie algebras, that do not lie in the enveloping algebra (are not polynomials in the generators) and their possible use in representation theory and physics. We have constructed such invariants for all subalgebras of $LO(4,1)$ for which they exist.

(iv) Diagrams characterizing subgroup reductions of the type shown in Fig. 8 have been used previously[40] for the Lorentz group $O(3,1)$. Similar diagrams have been used to characterize coordinates in $O(n)$ and $O(n,1)$ spaces (the "method of trees")[59,60]

IV. Separable coordinate systems

Let us consider the upper sheet of the two-sheeted hyperboloid

$$x_0^2 - x_1^2 - x_2^2 - x_3^2 - x_4^2 = 1. \tag{51}$$

If we consider a space of scalar functions $\psi(x)$ defined on this hyperboloid and require that a set of such functions transforms irreducibly under the group $O(4,1)$, then $\psi(x)$ must be eigenfunctions of the two Casimir operators of $O(4,1)$. However, the fourth-order Casimir operator is identically zero on such a space and the second order one reduces to the Laplace operator Δ on the hyperboloid (51).

We can now choose a basis by requiring that the functions $\psi(x)$ be eigenfunctions of Δ and of one of the complete sets of commuting operators, corresponding to one of the group reductions discussed above and represented in Fig. 8. It is interesting to note that to each subgroup reduction there corresponds a system of coordinates for which all the equations separate. For future convenience we write out these coordinate systems. It would be quite simple to present the Laplace operator in each system and also the eigenfunctions, but we do not do this here.

A. Reduction $O(4,1) \supset O(4)$

$$x_0 = \cosh a, \quad x_i = \sinh a \tilde{x}_i, \quad 0 \leq a < \infty, \quad i = 1, \ldots, 4,$$
$$\tilde{x}_1^2 + \tilde{x}_2^2 + \tilde{x}_3^2 + \tilde{x}_4^2 = 1. \tag{52}$$

On the sphere $\tilde{x}_i$ we then either introduce spherical coordinates [reduction to $O(3) \supset O(2)$], or cylindrical coordinates [reduction to $O(2) \oplus O(2)$]. For a complete discussion of all subgroup and nonsubgroup type coordinates on an O(4) sphere, see Ref. 61.

B. Reduction $O(4,1) \supset O(3,1)$

$$x_\mu = \sinh a \tilde{x}_\mu, \quad x_4 = \sinh a, \quad -\infty < a < \infty, \tag{53}$$
$$\mu = 0, 1, 2, 3, \quad \tilde{x}_0^2 - \tilde{x}_1^2 - \tilde{x}_2^2 - \tilde{x}_3^2 = 1.$$

On the $O(3,1)$ hyperboloid $\tilde{x}_\mu$ we introduce one of the seven types of subgroup coordinates, discussed earlier.[40,62] These are spherical coordinates for $O(3,1) \supset O(3)$, hyperbolic of three types for $O(3,1) \supset O(2,1)$, horospheric of two types for $O(3,1) \supset E(2)$, and cylindrical for $O(3,1) \supset O(2) \otimes O(1,1)$. Olevskiĭ also lists 27 nonsubgroup type coordinates for the $O(3,1)$ hyperboloid.

C. Reduction $O(4,1) \supset E(3)$

$$x_0 = \cosh\gamma + \tfrac{1}{2}(x^2 + y^2 + z^2) e^{-\gamma},$$
$$x_4 = \sinh\gamma + \tfrac{1}{2}(x^2 + y^2 + z^2) e^{-\gamma}, \tag{54}$$
$$x_1 = e^{-\gamma} x, \quad x_2 = e^{-\gamma} y, \quad x_3 = e^{-\gamma} z, \quad -\infty < \gamma < \infty.$$

On the Euclidean space x, y, z we can use spherical [$E(3) \supset O(3)$], cylindrical [$E(3) \supset E(2) \supset O(2)$] or Cartesian [$E(3) \supset E(2) \supset T(1)$] coordinates. All 11 subgroup and nonsubgroup type separable coordinates in E(3) space are discussed in the literature.[63,64]

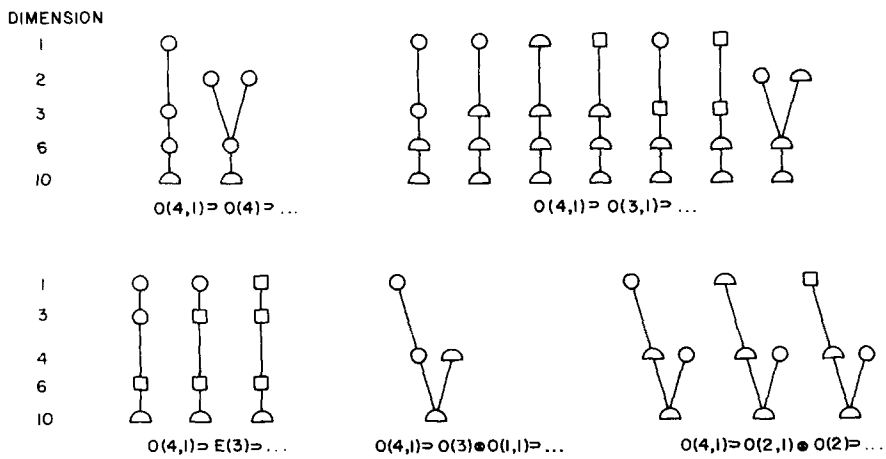

FIG. 8. Chains of subgroups of O(4, 1) providing quantum numbers and separable coordinate systems. A semicircle indicates an $O(p, 1)$ group, a circle an $O(n)$ group and a square an $E(n)$ group.

D. Reduction $O(4,1) \supset O(3) \oplus O(1,1)$

$$x_0 = \cosh a \cosh b, \quad x_2 = \sinh a \sin\theta \cos\phi,$$
$$x_1 = \cosh a \sinh b, \quad x_3 = \sinh a \sin\theta \sin\phi, \tag{55}$$
$$x_4 = \sinh a \cos\theta,$$
$$0 \leq a < \infty, \quad -\infty < b < \infty, \quad 0 \leq \theta \leq \pi, \quad 0 \leq \phi < 2\pi.$$

E. Reduction $O(4,1) \supset O(2,1) \oplus O(2)$

$$x_k = \cosh a \tilde{x}_k, \quad x_3 = \sinh a \cos\phi,$$
$$x_4 = \sinh a \sin\phi, \tag{56}$$
$$k = 0, 1, 2, \ \tilde{x}_0^2 - \tilde{x}_1^2 - \tilde{x}_2^2 = 1, \ 0 \leq a < \infty, \quad 0 \leq \phi < 2\pi.$$

On the $O(2,1)$ sphere $\tilde{x}_k$ we introduce spherical $[O(2,1) \supset O(2)]$, hyperbolic $[O(2,1) \supset O(1,1)]$ or horospheric $[O(2,1) \supset T(1)]$ coordinates.

The connection between Lie theory and the separation of variables has received a lot of attention in the literature.[40–43,59–61,65] The results of this section, aside from listing all subgroup type separable coordinates, suggest a recursive method for introducing separable coordinates in arbitrary $O(p,1)$ and more generally $O(p,q)$ spaces.

4. CONCLUSIONS

The main result of this paper is that we have provided a complete classification of the continuous subgroups of the de Sitter group $O(4,1)$. Thus, we have shown that $O(4,1)$ has four maximal subgroups, namely $O(4)$, $O(3,1)$, $D \square E(3)$ and $O(2) \otimes O(2,1)$. All continuous subgroups of these maximal subgroups were classified into conjugacy classes, where conjugacy was considered under the maximal subgroup, under $SO_0(4,1)$ [the connected component of $O(4,1)$], under $SO(4,1)$ and under $O(4,1)$. The results are summarized in Fig. 1–6. In particular representatives of all $O(4,1)$ conjugacy classes of subalgebras of $LO(4,1)$ are given on Fig. 6 which also shows their mutual inclusions.

In Sec. 3 we have found the invariants of all subalgebras of $LO(4,1)$ (if such exist), both Casimir operators, i.e., polynomial invariants, lying in the center of the enveloping algebra, and also nonpolynomial invariants. These were used to present different possible sets of commuting operators, providing quantum numbers for an elementary physical system, described by an irreducible unitary representation of $O(4,1)$. A lattice of subgroups with Casimir operators is given in Fig. 7. A graphical representation of all different chains of such subgroups is shown in Fig. 8. We have also given a list of all "subgroup type" systems of coordinates, allowing the separation of variables in the Laplace equation on the hyperboloid $x_0^2 - x_1^2 - x_2^2 - x_3^2 - x_4^2 = 1$.

It should be mentioned that the subgroup structure of $O(4,1)$ is relatively simple—much more so than that of the other groups of immediate physical interest, like the Poincaré group,[38] the similitude group,[39] the $O(3,2)$ de Sitter group, or the conformal group of space–time. These last two groups will be the subject of subsequent publications.

We have not gone deeply into any applications, however, the physical context in which the present results should be useful was discussed in the Introduction, and we plan to return to this separately.

*Work partially supported by NATO.
†Permanent address: Department of Mathematics, Ohio State University, Columbus, Ohio.

[1] W. de Sitter, Month. Not. 78, 3 (1917).
[2] G.E. Lemaître, J. Math. and Phys. (MIT) 4, 188 (1925).
[3] H.P. Robertson, Phil. Mag. 5, 835 (1928).
[4] P.A.M. Dirac, Ann. Math. 36, 657 (1935).
[5] F. Gürsey, in *Group Theoretical Concepts and Methods in Elementary Particle Physics*, edited by F. Gürsey (Gordon and Breach, New York, 1964).
[6] F. Gürsey and T.D. Lee, Proc. Nat. Acad. Sci. USA 49, 179 (1963).
[7] T.O. Philips and E.P. Wigner, in *Group Theory and Applications*, Vol. 1, edited by E.M. Loebl (Academic, New York, 1968).
[8] K.C. Hannabuss, Proc. Camb. Phil. Soc. 70, 283 (1971).
[9] M.M. Bakri, J. Math. Phys. 10, 298 (1969); 11, 2027 (1970).
[10] P. Roman and J.J. Aghassi, Phys. Lett. 14, 68 (1965); Nuovo Cimento 36, 1062 (1965); 37, 354 (1965); 38, 1092 (1965); J. Math. Phys. 7, 1273 (1966).
[11] O. Nachtmann, Comm. Math. Phys. 6, 1 (1967).
[12] L.H. Thomas, Ann. Math. 42, 113 (1941).
[13] T.D. Newton, Ann. Math. 51, 730 (1950).
[14] J. Dixmier, Bull. Soc. Math. France 89, 9 (1961).
[15] L. O'Raifeartaigh, Phys. Rev. 139, 1052 (1965).
[16] W. Tait and J.F. Cornwell, J. Math. Phys. 12, 1651 (1971).
[17] A.J. Bracken and H.A. Cohen, J. Math. Phys. 10, 2024 (1969).
[18] L.L. Foldy and S.A. Wouthuysen, Phys. Rev. 78, 29 (1950).
[19] A. Inomata, in *De Sitter and Conformal Groups and Their Applications*, Lectures in Theoretical Physics, edited by A.O. Barut and W.E. Brittin (Colorado Assoc. U.P., Boulder, Colorado, 1971), Vol. 13.
[20] A.O. Barut and A. Bohm, J. Math. Phys. 11, 2938 (1970).
[21] E. Inonu and E.P. Wigner, Proc. Nat. Acad. Sci. USA 39, 510 (1953).
[22] E. Saletan, J. Math. Phys. 2, 1 (1961).
[23] H. Bacry and J.M. Lévy-Leblond, J. Math. Phys. 9, 1605 (1968).
[24] J.R. Derome and J.G. Dubois, Nuovo Cimento B 9, 351 (1972); J.G. Dubois, Can. J. Phys. 51, 1757 (1973), Nuovo Cimento B 15, 1 (1973).
[25] M.Y. Han, Nuovo Cimento B 42, 367 (1966).
[26] H. Bacry, Nuovo Cimento A 41, 222 (1966).
[27] I.A. Malkin and V.I. Man'ko, Yad. Fiz. 3, 372 (1966) [Sov. J. Nucl. Phys. 3, 267 (1966)].
[28] Y. Dothan, M. Gell-Mann, and Y. Ne'eman, Phys. Lett. 17, 148 (1965).
[29] M. Mukunda, L. O'Raifeartaigh, and E.C.G. Sudarshan, Phys. Rev. Lett. 15, 1041 (1965); Phys. Lett. 19, 322 (1965).
[30] R.H. Pratt and T.F. Jordan, Phys. Rev. 148, 1276 (1966).
[31] A.O. Barut and H. Kleinert, Phys. Rev. 156, 1541 (1967); 157, 1180 (1967).
[32] M.J. Englefield, *Group Theory and the Coulomb Problem* (Wiley Interscience, New York, 1972).
[33] M. Bander and C. Itzykson, Rev. Mod. Phys. 38, 330, 346 (1966).
[34] M.J. Cunningham, J. Math. Phys. 13, 33, 1108 (1972).
[35] R. Bogdanović and M.A. Whitehead, J. Math. Phys. 16, 400 (1975).
[36] J. Patera, P. Winternitz, and H. Zassenhaus, J. Math. Phys. 15, 1378 (1974).
[37] J. Patera, P. Winternitz, and H. Zassenhaus, J. Math. Phys. 15, 1932 (1974).
[38] J. Patera, P. Winternitz, and H. Zassenhaus, J. Math. Phys. 16, 1597 (1975).
[39] J. Patera, W. Winternitz, and H. Zassenhaus, J. Math. Phys. 16, 1613 (1975).

[40]P. Winternitz and I. Friš, Yad. Fiz. **1**, 889 (1965) [Sov. J. Nucl. Phys. **1**, 636 (1965)].
[41]P. Winternitz, I. Lukač, and Ya.A. Smorodinskiĭ, Yad. Fiz. **7**, 192 (1968) [Sov. J. Nucl. Phys. **7**, 139 (1968)].
[42]J. Patera and P. Winternitz, J. Math. Phys. **14**, 1130 (1973).
[43]E.G. Kalnins and W. Miller Jr., J. Math. Phys. **15**, 1025, 1263, 1728 (1974).
[44]G.W. Mackey, The Theory of Group Representations (Univ. of Chicago Lecture Notes, 1955).
[45]S. Helgason, *Differential Geometry and Symmetric Spaces* (Academic, New York, 1962).
[46]D. Finkelstein, Phys. Rev. **100**, 924 (1955).
[47]A. Kihlberg, Arkiv Fysik **34**, 307 (1967).
[48]H. Bacry and A. Kihlberg, J. Math. Phys. **10**, 2132 (1969).
[49]H. Zassenhaus, Abhandl Math. Sem. Hansisch. Univ. **12**, 289 (1938).
[50]B. Meyer, Can. J. Math. **6**, 155 (1953).
[51]M.S. Raghunathan, *Discrete Subgroups of Lie Groups* (Springer, New York, 1972).
[52]E.G. Kalnins, J. Patera, R.T. Sharp, and P. Winternitz, "Elementary Particle Reactions and the Lorentz and Galilei Groups," in *Group Theory and its Applications, Vol.* 3, edited by E.M. Loebl (Academic, New York, 1975).
[53]J. Patera, P. Winternitz, and H. Zassenhaus, "Continuous Subgroups of the Fundamental Groups of Physics. 3. The de Sitter Groups," Preprint CRM, 1975, to be published).
[54]E.G. Kalnins, J. Patera, R.T. Sharp, and P. Winternitz, Phys. Rev. D **8**, 2552 (1973).
[55]E. Goursat, Ann. Sci. l'Ecole Supérieure **6** (3) 9 (1889).
[56]E.L. Ince, *Ordinary Differential Equations* (Dover, New York, 1944).
[57]B. Judd, W. Miller Jr., J. Patera, and P. Winternitz, J. Math. Phys. **15**, 1787 (1974); M. Moshinsky, J. Patera, R.T. Sharp, and P. Winternitz. Ann. Phys. (N.Y.) **95**, 139 (1975).
[58]C.P. Boyer and K.B. Wolf, J. Math. Phys. **15**, 560 (1974).
[59]N.Ya. Vilenkin, Mat. Sbornik (N.S.) **68**, (110), 432 (1965).
[60]N.Ya. Vilenkin, G.I. Kuznetsov, and Ya.A. Smorodinskiĭ, Yad. Fiz. **2**, 906 (1965). [Sov. J. Nucl. Phys. **2**, 645 (1966)].
[61]E.G. Kalnins, W. Miller, Jr., and P. Winternitz, SIAM J. Appl. Math. **30**, No. 4 (June 1976).
[62]M.P. Olevskiĭ, Mat. Sb. **27** (69), 379 (1950).
[63]Ph.M. Morse and H. Feshbach, *Methods of Theoretical Physics, Part* 1 (McGraw-Hill, New York, 1953).
[64]P. Moon and D.E. Spencer, *Field Theory Handbook* (Springer-Verlag, Berlin, 1961).
[65]N.Y. Vilenkin and Ya.A. Smorodinskiĭ, Zh. Eksp. Teor. Fiz. **46**, 1793 (1964) [Sov. Phys. JETP **19**, 1209 (1964)].

n-body scattering into cones with long-range time-dependent potentials

Joseph H. Hendrickson

4872 E. Gettysburg, Fresno, California 93726
(Received 27 March 1975; revised manuscript received 16 September 1975)

Modified wave operators are shown to exist for n particles with long-range, time-dependent potentials. Bound states must be assumed to be "slightly better" than L^2. As an application of these temporally inhomogeneous modified wave operators, results of Dollard on scattering into cones are extended to this context.

The purpose of this paper is to establish the existence of modified wave operators for n particles with long-range, time-dependent potentials. Results of Dollard[1] on scattering into cones are extended to this context. If one considers the Dollard modified wave operator,

$$W_{\pm,D} = \text{s-}\lim_{t\to\pm\infty} \exp(itH)\exp(-itH_0 - i\int_0^t V(p\tau)\,d\tau),$$

where $p = -i\nabla$, and $H_0 = -\Delta/2$, one notes that

$$\bar{U}_0(t,0) = \exp(-itH_0 - i\int_0^t V(p\tau)\,d\tau),$$

where $\bar{U}_0(t,s)$ is the evolution operator for the time-dependent Hamiltonian $H_0 + V(pt)$. From this perspective on the "anomalous term" $-i\int_0^t V(pt)\,d\tau$, the meaning of the modified wave operator is clear. The existence of the standard wave operators for a short-range potential $V^S(x)$ asserts that the dynamics of $H_0 + V^S(x)$ are asymptotically approximated by those of the simpler free Hamiltonian H_0. Similarly, the existence of the Dollard modified wave operator for a long-range potential $V^L(x)$ asserts that, while the asymptotic dynamics are too complex to be approximated by the simple free motion, they may be approximated by the dynamics of $H_0 + V^L(pt)$. While $H_0 + V^L(pt)$ is time-dependent and, therefore, looks complicated, it is generally easier to work with than the original Hamiltonian $H_0 + V^L(x)$ since it has no "x" dependence and, therefore, preserves momentum and commutes with functions of "p." The natural way in which this time-dependent Hamiltonian arises suggests that time-dependent Hamiltonians form the proper context for the study of modified wave operators.

In an earlier paper,[2] the author generalized the long-range, modified wave operator existence theorem of Alsholm and Kato[3] to include time-dependent Hamiltonians: the temporally inhomogeneous modified wave operators

$$W_{\pm,D}(s) = \text{s-}\lim_{t\to\pm\infty} U(s,t)\bar{U}_0(t,s)$$

were shown to exist, where $U(t,s)$ and $\bar{U}_0(t,s)$ are the evolution operators for $H(t) = H_0 + V(t,x)$ and $H_0 + V(t,pt)$ respectively, and where uniform growth conditions on $V(t,x)$ were assumed. In the present paper we wish to illustrate the generality and validity of temporally inhomogeneous modified wave operators by generalizing the existence theorem to n bodies and by showing that they behave nicely, conforming to Dollard's work on scattering into cones.

Our notation closely follows that of Dollard. Because of the confusing amount of notation, we have made the assumptions that all particles have mass 1 and that there is no static potential. Neither assumption affects the argument in Theorem 1. In the presence of static potentials certain limits in Theorem 2 would have to be replaced by limits in the Cesaro sense, cf. Dollard.[1] Considering n particles in $\mathbb{R}^3$ gives the $3n$-dimensional coordinates $x = (x_1, \ldots, x_n) \in \mathbb{R}^{3n}$, each $x_i \in \mathbb{R}^3$. If $\Gamma = \{\Gamma_1, \ldots, \Gamma_m\}$ is a partition of the n particles into fragments of $r_l + 1$ particles each, $l = 1, \ldots, m$, we define the following internal coordinates:

$$y_l = \frac{1}{r_l + 1}\sum_{j\in\Gamma_l} x_j, \quad \text{(center of mass coordinates)}$$

$$z_l = (z_l^1, \ldots, z_l^{r_l})$$

$$= (x_{j_1} - x_{j_2}, x_{j_2} - x_{j_3}, \ldots, x_{j_{r_l}} - x_{j_{r_l+1}}), \quad \text{(relative coordinates).}$$

Let $\phi_\alpha(z_\alpha) = \phi_1(z_1)\phi_2(z_2)\cdots\phi_m(z_m)$ be bound states for the fragments. Then $\alpha = (\Gamma, \phi_\alpha)$ is a channel. We write $y_\alpha = (y_1, \ldots, y_m)$ and $z_\alpha = (z_1, \ldots, z_m)$ and omit the α where the context is clear. If E_l is the energy of $\phi_l(\cdot)$, we write $E_\alpha = \sum_{l=1}^m E_l$ for the energy of ϕ. Note:

$x \mapsto (y_\alpha, z_\alpha)$ has Jacobian 1,

$$\frac{\partial}{\partial y_l} = \sum_{j\in\Gamma_l}\frac{\partial}{\partial x_j} \equiv ip_l,$$

write $p_\alpha = (p_1 \ldots, p_m) = -i\nabla_{y_\alpha}$

$$H_0 = \frac{-\Delta}{2} = H_{y_\alpha} + H_{z_\alpha}$$

where

$$H_{y_\alpha} = -\Delta y_\alpha/2,$$

$$H_{z_\alpha} = \sum_{l=1}^m H_{z_l} \text{ and each } H_{z_l} \text{ is a function only of the partials with respect to } z_l.$$

Given a function $F(\cdot)$, we shall often write $F(x) = F(y,z)$. We also shall often write "$j \in l$" to mean "the jth particle is in Γ_l."

Naturally, restrictions must be imposed on the potential $V(t,x)$. To begin with, it must be assumed that $U(t,s)$, the evolution operator for $H(t) = H_0 + V(t,x)$, exists. For relevant conditions on $V(t,x)$ see Goldstein

and Monlezun[4] or Kato.[5] Since there is no static potential, $V(t,x)$ is of the form

$$V(t,x)=\sum_{\substack{i\in I\subseteq \bar{I}\ni j\\ i\neq j}} V_{ij}(t,x_i-x_j).$$

Each $V_{ij}(t,x)$ is the sum of a short-range and a long-range potential where growth conditions are imposed on the long-range potential; that is,

(AI). Assume $V_{ij}(t,x_i-x_j)=V^S_{ij}(t,x_i-x_j)+V^L_{ij}(t,x_i-x_j)$ for all $i,j=1,\ldots,n$ and

(a) There exist positive constants c,ϵ,β,γ such that $1\geq\beta>\frac{1}{2}$ and $\gamma>(1-\beta)^2\beta^{-1}$ and such that the following hold for all $i,j=1,\ldots,n$:

(b) $|V^S_{ij}(t,x_i-x_j)|\leq c(1+|x_i-x_j|)^{-1-\epsilon}$,

(c) $D_t^{\xi_1}D_x^{\xi_2}V^L_{ij}(t,x_i-x_j)\in L^1_{loc}(\mathbb{R}^{3n+1})$ for $|\xi_1|=0,1$ and $|\xi_2|=0,1,2,3$ where the derivatives are taken in the sense of distribution theory,

(d) $|D_x^{\xi}V^L_{ij}(t,x_i-x_j)|$

$$\leq\begin{cases} c(1+|x_i-x_j|)^{-1-\beta}, & |\xi|=1,\\ c(1+|x_i-x_j|)^{-2-\gamma}, & |\xi|=2,3,\end{cases}$$

(e) $|D_tD_x^{\xi}V^L_{ij}(t,x_i-x_j)|$

$$\leq\begin{cases} c(1+|t|)^{-1}(1+|x_i-x_j|)^{-1-\beta}, & |\xi|=1,\\ c(1+|t|)^{-1}(1+|x_i-x_j|)^{-2-\gamma}, & |\xi|=2,3.\end{cases}$$

Corresponding to the channel α, we partition $V(t,x)$ as follows:

$$V(t,x)=V^S_\alpha(t,x)+V^L_\alpha(t,x)+V^I_\alpha(t,x),\ \text{where}$$

$$V^S_\alpha(t,x)=\sum_{i\in I\not\subseteq \bar{I}\ni j} V^S_{ij}(t,x_i-x_j),$$

$$V^L_\alpha(t,x)=\sum_{i\in I\not\subseteq \bar{I}\ni j} V^L_{ij}(t,x_i-x_j),$$

$$V^I_\alpha(t,x)=\sum_{I=1}^{m}\sum_{i,j\in I} V_{ij}(t,x_i-x_j).$$

$V^S_\alpha(t,x)$ and $V^L_\alpha(t,x)$ are the short and long-range potentials between fragments while $V^I(t,x)$ is the internal potential within the fragments. The assumption that we asymptotically have stable fragments with energy E_α means that the interval potential must stabilize at infinity and $H_{z_\alpha}+V^I_\alpha(t,x)$ must converge in some sense to E_α.

(A II) Assume that the L^2-norm of $(H_{z_\alpha}+V^I_\alpha(t,x)-E_\alpha)\phi_\alpha(z)$ as a function of z (i.e., $\|[H_{z_\alpha}+V^I_\alpha(t,x)-E_\alpha]\phi_\alpha(z_\alpha)\|_{L^2(\mathbb{R}^{3n-m})}$) is integrable as a function of t.

This is now sufficient notation to define the channel operator $H_\alpha(t)=H_{y_\alpha}+V^L_\alpha(t,p_\alpha t,0)+E_\alpha$. The short-range potential between fragments does not appear since it is asymptotically negligible. The long-range potential between fragments is written as a function of the relative coordinates (y_α,z_α). As in the one body case, y_α is replaced by $p_\alpha t=-i\nabla_{y_\alpha}t$; the internal coordinates z_α have been ignored since the distance between fragments could be expected to soon overwhelm the distances within fragments. To insure that this happens quickly enough, we assume:

(A III) $z_\alpha\mapsto\phi_\alpha(z_\alpha)|z_\alpha|$ is square integrable. This is similar to, though stronger than, Dollard's condition[1,6,7] and appears to be reasonable because of "the usual exponential dampin of bound-state wavefunctions."[7]

Theorem 1: Let P_α project onto the closed subspace D_α generated by $\{f(y_\alpha)\phi_\alpha(z_\alpha)|f\in L^2(\mathbb{R}^m)\}$. Assume that the evolution operator for $H(t)$ exists. Assume (A I), (A II), and (A III). Then the temporally inhomogeneous modified channel operators,

$$W^\alpha_{\pm,D}(s)=\text{s-}\lim_{t\to\pm\infty}U(s,t)U_\alpha(t,s)P_\alpha$$

exist, where $U(t,s)$ and $U_\alpha(t,s)$ are the evolution operators for $H(t)=H_0+V(t,x)$ and $H_\alpha(t)$, respectively.

Before proving Theorem 1, we state the following four propositions.

Proposition 1: Let $F:\mathbb{R}^m\to\mathbb{R}$ be measurable. Then the following identity of operators holds,

$$\exp(it|p_\alpha|^2/2)F(y_\alpha)\exp(-i|p_\alpha|^2/2)$$
$$=\exp(-i|y_\alpha|^2/2t)F(p_\alpha t)\exp(i|y_\alpha|^2/2t).$$

Proposition 2 : $U_\alpha(t,s)$ commutes with P_α.

Proposition 3 : If $W^\alpha_{\pm,D}(s)$ exists for one $s\in\mathbb{R}$ then $W^\alpha_{\pm,D}(r)$ exists for all $r\in\mathbb{R}$ and $W^\alpha_{\pm,D}(r)=U(r,s)W^\alpha_{\pm,D}(s)\times U_\alpha(s,r)$.

Proposition 4:

$$|U_\alpha(t,s)f(tw_1,\ldots,tw_m)|\leq ct^{-3m/2}(1+w_1^2)^{-1}\cdots(1+w_m^2)^{-1}.$$

For the proof Proposition 1, see Hendrickson[8] or Alsholm.[9] The proof of Proposition 2 is evident. The proof of Proposition 3 follows from the one-body case[2,8] and Proposition 2.

Proposition 4 is derived in the same way as Dollard's[7] equation (66) and is based on the identity:

$$U(t,s)f(y)=\int dy'\prod_{l=1}^{m}[2\pi i(t-s)]^{-3/2}\exp[i(y_l-y_l')^2/2t]$$
$$\times\exp[-i\int_s^t V(\tau,p\tau,0)\,d\tau]f(y')$$

using integration by parts and (AId).

Proof of Theorem 1: The complications of the proof that are due to the time dependence are handled in the same way as in the time-dependent one-body case.[2,8] Those complications due to the n-bodies are handles as in Dollard's time-independent n-body proof.[7] Therefore, the proof given here will be brief. The channel α is constant throughout and will often be omitted as a subscript. By Proposition 3, we may assume $s=0$. It suffices to show convergence on the dense set where $\hat{f}$, the Fourier transform of f, is C^∞ with compact support bounded away from the lines $p_l=p_{\bar{l}}$ for $l\neq\bar{l}$. Write $X_t=\int_0^t V^L(\tau,p\tau,0)\,d\tau$. By the standard Cook reduction,[2]

it suffices to show the integrability of

$$\begin{aligned} a(t) &= \| [H(t) - H_\alpha(t)]U_\alpha(t,s)f(y)\phi(z) \| \\ &\le \| V_\alpha^s(t,x)U_\alpha(t,s)f(y)\phi(x) \| \\ &\quad + \| [V_\alpha^L(t,y,z) - V_\alpha^L(t,y,0)]U_\alpha(t,s)f(y)\phi(z) \| \\ &\quad + \| [V_\alpha^L(t,y,0) - V_\alpha^L(t,pt,0)]U_\alpha(t,s)f(y)\phi(z) \| \\ &\quad + \| [H_z + V^I(t,x) - E_\alpha]U_\alpha(t,s)f(y)\phi(z) \| \\ &\equiv a_1(t) + a_2(t) + a_3(t) + a_4(t). \end{aligned}$$

For $j \in l$, $x_j = y_l + \sum_{k=1}^{r_l} \lambda_j^k z_l^k$ for some λ_j^k. If $i,j \in l$, then $x_i - x_j$ is a function only of z and then $a_4 = \| U_\alpha(t,s)f(y) \|_y \cdot \| [H_z + V_\alpha^I(t,x) - E_\alpha]\phi(z) \|_z$ is integrable by (AII) where $\| \ \|_y$ and $\| \ \|_z$ are the L^2-norms in $\mathbf{R}^m$ and $\mathbf{R}^{3n-m}$ with respect to y and z.

$$\begin{aligned} a_3(t) &= \| \phi(z) \|_z \| [V_\alpha^L(t,y,0) \\ &\quad - V_\alpha^L(t,pt,0)]U_\alpha(t,s)f(y) \|_y \end{aligned}$$

is in the form of the one-body problem and is bounded using Proposition 2.[2] $a_1(t)$ will be bounded as $a_2(t)$ below.

In bounding $a_2(t)$ we use the notation $w_l = y_l/t$, $l = 1, \ldots, m$, and, for $i \in l < \bar{l} \ni j$, define

$$h_{i,j}(z) \equiv \sum_{k=1}^{r_l} \lambda_i^k z_l^k - \sum_{k=1}^{r_{\bar{l}}} \lambda_j^k z_{\bar{l}}^k = x_i - x_j - y_l + y_{\bar{l}}.$$

Then

$$\begin{aligned} a_2(t) &\le \sup_{\lambda \in [0,1]} \| \int_1^0 |\nabla_x V(t, y, \lambda z)| \cdot |z| d_\lambda U_\alpha(t,s)f(y)\phi(z) \| \\ &\le c \sum_{i \in l < \bar{l} \ni j} \sup_{\lambda \in [0,1]} \| (1 + |y_l - y_{\bar{l}} + h_{i,j}(\lambda z)|)^{-1-\beta} \\ &\quad \times |z| U_\alpha(t,s)f(y)\phi(z) \| \\ &\equiv c \sum_{i \in l < \bar{l} \ni j} a(i,j,t), \end{aligned}$$

for some constant c. Taking for convenience the case $l = 1$ and letting $w_0 = y_1 - y_{\bar{l}}$ for some fixed $\bar{l}$, we get

$$\begin{aligned} &[a(i,j,t)]^2 \\ &\le \sup_{\lambda \in [0,1]} \| |z| (1 + w_0 + h_{ij}(\lambda z))^{-1-\beta} \\ &\quad \times U_\alpha(t,s)f(y)\phi(z) \|^2 \\ &= \sup_{\lambda \in [0,1]} \int \Big| |z| (1 + |w_0 + h_{ij}(\lambda z)|)^{-1-\beta} \\ &\quad \times U_\alpha(t,s)f(w_0, y_2, \ldots, y_m)\phi(z) \Big|^2 dw_0 dy_2 \cdots dy_m dz \\ &\le \sup_{\lambda \in [0,1]} t^{3(m-1)} \int \Big| |z| (1 + |w_0 + h_{ij}(\lambda z)|)^{-1-\beta} \\ &\quad \times U_\alpha(t,s)f(w_0, tw_2, \ldots, tw_m)\phi(z) \Big|^2 dw_0 dw_2 \cdots dw_m dz \\ &\le \sup_{\lambda \in [0,1]} t^{3(m-1)} \int dz\, |z|^2 |\phi(z)|^2 \Big[\sup_{w_0 \in \mathbf{R}^3} \Big| \int U_\alpha(t,s) \\ &\quad \times f(w_0, tw_2, \ldots, tw_m) \Big|^2 dw_2 \cdots dw_m \Big] \cdot \int (1 + |w_0 \\ &\quad + h_{ij}(\lambda z)|)^{-2-2\beta} dw_0 \\ &\le ct^{-3} \end{aligned}$$

using (AI), Proposition 4, and $\beta > \frac{1}{2}$. ∎

Now that we know that at least some temporally inhomogeneous modified channel operators exist, we may reasonably assume their existence and see if they behave in a reasonable manner. Suppose that in a scattering experiment one "sent in" n particles with asymptotic energy $H_\beta(t)$ along channel β; they interact somehow, and then leave with asymptotic energy $H_\alpha(t)$ along channel α. If $f_\beta(s, \ )$ represents the "semifree" state governed by $H_\beta(t)$ at time s, then the actual state at time s is given by $u(s, \ ) = W_-^\beta(s)f_\beta(s, \ )$, and at an arbitrary time t is given by $u(t, \ ) = U(t,s)W_-^\beta(s)f_\beta(s, \ )$. If $C_1, \ldots, C_n \subseteq \mathbf{R}^3$ are cones with vertices at the origin, then the probability that one would asymptotically find the ith particle in C_i for $i = 1, \ldots, n$ is

$$\begin{aligned} P(f_\beta) &= \lim_{t \to -\infty} \int_{C_1 \times \cdots \times C_n} |U(t,s)W_-^\beta(s)f_\beta(s,x)| dx \\ &= \lim_{t \to -\infty} \int_{C_1 \times \cdots \times C_n} |U(t,r)W_-^\beta(r)f_\beta(r,x)| dx. \end{aligned} \tag{1}$$

The second equality follows from Proposition 3 and is given to illustrate the independence of $P(f_\beta)$ on time.

$P(f_\beta)$ is what we wish to look at. What result should one expect? Consider first a free particle. Intuitively, the probability that a free particle is asymptotically in a cone should be the same as the probability that its momentum is in the cone. (To see this just draw a cone and the straight-line path of the particle in the direction of its constant momentum.) Switching to a single, nonfree particle, it follows that the probability of an incoming particle ending up in a given cone is determined by the probability that the momentum of the outgoing particle is in the cone. In generalizing to the n-body case, where one visualizes each fragment clustered around its center of mass, one should only need to look at the probability of the center-of-mass coordinates being in the intersection of all the cones associated with the particles of the fragment. This is what Dollard[1] showed in his paper. Since the semifree Hamiltonian $H_\beta(t)$ preserves momentum, the same analysis should be valid in the present context. This is what we wish to check. In what follows, s is held constant and $f_\beta(s, \ )$ is abreviated as f_β.

Let $f_\beta \in D_\beta$. Define[1] $g_{\alpha\beta}^s$ by

$$S_{\alpha\beta}(s)f_\beta(x) \equiv W_{+,D}^\alpha(s)^{-1} W_{-,D}^\beta(s)f_\beta(x) = g_{\alpha\beta}^s(y)\phi_\alpha(z), \tag{2}$$

where $\alpha = (\Gamma, \phi)$ and $\Gamma = \{\Gamma_1, \ldots, \Gamma_m\}$. Define $I_{\alpha l} = \cap_{j \in \Gamma_l} C_j$, $l = 1, \ldots, m$. Then Theorem 2 follows.

Theorem 2: Assume $\lim_{t \to \pm\infty} 1/t^{|\xi|} \int_0^t D_p^\xi V_{ij}^L(\tau, p\tau, 0)\, d\tau = 0$ uniformly in p on compact sets for $1 \le |\xi| \le 3n$; assume there is no static potential, and assume $W_{\pm,D}^\alpha(s)$ exists for fixed s and for all α, and are asymptotically complete. Then

$$P(f_\beta) = \sum_\alpha \int_{I_{\alpha l_1} \times \cdots \times I_{\alpha l_m}} |\widehat{g_{\alpha\beta}^s}(p_1, \ldots, p_n)| dp_1 \cdots dp_n.$$

We first make some comments on the restrictions in the theorem. The first assumption is not very strong since archetypical examples of long-range, time-dependent potentials have the basic form $1/(1 + |tx|)$ or $c(t)/(1 + |x|)$ where $c(\cdot)$ is the characteristic function of some finite interval. Also the growth requirements of the existence theorems are much stronger than this re-

striction, although for fewer derivatives. Static potential is assumed zero to avoid bound states which are not eventually in any cone. By the term "asymptotic completeness" it is meant that $W^{\alpha}_{\pm,D}(s)$ exist for all channels α and fixed s (which implies for all s); that if $R^{\pm}_{\alpha}(s)$ is the range of $W^{\alpha}_{\pm,D}(s)$, then $R^{+}_{\alpha}(s)=R^{-}_{\alpha}(s)$ for fixed s (which implies for all s by Proposition 5 below); and that $R^{\pm}(s)\equiv\cup_{\alpha}R^{\pm}_{\alpha}(s)$ is the orthogonal complement of the bound states of the static potential. Thus, zero static potential will imply $R^{\pm}(s)=L^2(\mathbb{R}^{3n})$.

Proposition 5: If $R^{+}_{\alpha}(s)=R^{-}_{\alpha}(s)$ for fixed $s\in\mathbb{R}$, Then $R^{+}_{\alpha}(r)=R^{-}_{\alpha}(r)=U(r,s)R^{\pm}_{\alpha}(s)$ for all $r\in\mathbb{R}$.

Proof: $W^{\alpha}_{\pm,D}(s)=U(s,r)W^{\alpha}_{\pm,D}(r)U_{\alpha}(r,s)$ implies that $R^{\pm}_{\alpha}(s)\subseteq U(s,r)R^{\pm}_{\alpha}(r)$. Also

$$U(s,r)W^{\alpha}_{\pm,D}(r)=U(s,r)W^{\alpha}_{\pm,D}(r)U_{\alpha}(r,s)U_{\alpha}(s,r)$$
$$=W^{\alpha}_{\pm,D}(s)U_{\alpha}(s,r),\text{ implies that}$$
$$U(s,r)R^{\pm}_{\alpha}(r)\subseteq R^{\pm}_{\alpha}(s).\quad\blacksquare$$

The general outline of the proof of Theorem 2 is dictated by the following equality derived from Eqs. (1) and (2):

$$P(f_\beta)=\lim_{t\to\infty}\int_{C_1\times\cdots\times C_n}\Big|\sum_{\alpha}U_{\alpha}(t,s)S_{\alpha\beta}(s)\times f_\beta(x_1,\ldots,x_n)\Big|^2dx_1\cdots dx_n,\tag{3}$$

where f_β is the incoming state at time s. Again, to replace s by r, also replace f_β with $U(r,s)f_\beta$. The body of the proof is contained in a number of lemmas which evaluate the terms of the above series. Aside from modification of the lemmas to handle the additional term $V(t,pt,0)$, the main change from Dollard's proof is the elimination of the somewhat cumbersome "Q_t" and "C_t" transformations of Dollard by use of Proposition 1. We begin, unfortunately, with some more notation:

$$\chi_C(x)=\begin{cases}1, & \text{if } x\in C_1\times\cdots\times C_n,\\ 0, & \text{otherwise.}\end{cases}$$

$$\chi_{C_j}(x)=\begin{cases}1, & \text{if } x_j\in C_j,\\ 0, & \text{otherwise.}\end{cases}$$

$$\chi_{B_{\alpha l}}=\prod_{j\in\Gamma_l}\chi_{C_j}$$

$$\chi_{I_{\alpha l}}(r)=\begin{cases}1, & \text{if } r\in I_{\alpha l}\subseteq\mathbb{R}^3,\\ 0, & \text{otherwise.}\end{cases}$$

Lemma 1: Let $f_\pm=\sum_\alpha W^{\alpha}_{\pm}(s)f_\alpha$ with $f_\alpha\in D_\alpha$. Let $U_\pm(t,s)f_\pm=\sum_\alpha U_\alpha(t,s)f_\alpha$. Then

$$\lim_{t\to\pm\infty}\|U(t,s)f_\pm-U_\pm(t,s)f_\pm\|=0.$$

Proof: See Dollard.[1]

Lemma 2:

$$\text{s-}\lim_{t\to\pm\infty}\chi_{\beta_{\alpha l}}(x_1,\ldots,x_n)\Big|_{y_l=p_l t}=\chi_{I_{\alpha l}}(p_l)\quad\text{for } p_l\neq 0.$$

Proof: See Dollard.[1]

Lemma 3: If $A(t)$ is uniformly bounded and linear for all t, then

$$\text{s-}\lim_{t\to\pm\infty}A(t)\exp(-i|y_l|^2/2t)-\exp(-i|y_l|^2/2t)A(t)=0$$

Proof: Let $B(t)=\exp(-i|y_l|^2/2t)$. Then

$$\|A(t)B(t)-B(t)A(t)\|$$
$$=\|B(t)[B(t)^{-1}A(t)B(t)-A(t)]\|$$
$$\le\|B(t)\|[\|B(t)^{-1}A(t)B(t)-A(t)B(t)\|+\|A(t)B(t)-A(t)\|]$$
$$\le\|B(t)\|[\|B(t)^{-1}-1\|\cdot\|A(t)B(t)\|+\|A(t)\|\cdot\|B(t)-1\|]$$
$$\to 0\text{ as } t\to\pm\infty.$$

Notation: Let

$$X^t_s=\int_s^t V^L_\alpha(\tau,p\tau,0)\,d\tau.$$

Lemma 4:

$$\text{s-}\lim_{t\to\pm\infty}\exp(iX^t_s)\exp(i|y_l|^2/2t)\,\chi_{B_{\alpha l}}\Big|_{y_l=p_l t}\times\exp(-i|y_l|^2/2t)\exp(-iX^t_s)$$
$$=\chi_{I_{\alpha l}}(p_l).$$

Proof: Let $A(t)=\exp(iX^t_s)$ and $B(t)=\exp(i|y_l|^2/2t)$ and $\chi_B=\chi_{B_{\alpha\beta}}|_{y_l=p_l t}$. We use Lemma 2, Lemma 3, the fact that $A(t)$ and χ_B commute, and the fact that $\|\chi_B\|\le 1$.

$$\|A(t)B(t)\chi_B B(t)^{-1}A(t)^{-1}-\chi_{I_{\alpha\beta}}(p_l)\|$$
$$\le\|A(t)B(t)\chi_B B(t)^{-1}A(t)^{-1}-B(t)A(t)\chi_B B(t)^{-1}A(t)^{-1}\|$$
$$+\|B(t)A(t)\chi_B B(t)^{-1}A(t)^{-1}-B(t)A(t)\chi_B A(t)^{-1}B(t)^{-1}\|$$
$$+\|B(t)A(t)\chi_B A(t)^{-1}B(t)^{-1}-\chi_{I_{\alpha\beta}}(p_l)\|$$
$$\le\|A(t)B(t)-B(t)A(t)\|\,\|\chi_B B(t)^{-1}A(t)^{-1}\|$$
$$+\|B(t)A(t)\chi_B\|\,\|B(t)^{-1}A(t)^{-1}-A(t)^{-1}B(t)^{-1}\|$$
$$+\|B(t)\chi_B B(t)^{-1}-\chi_{I_{\alpha\beta}}(p_l)\|$$
$$\to 0\text{ as } t\to\pm\infty$$

Lemma 5:

$$\text{s-}\lim_{t\to\pm\infty}U_\alpha(s,t)\chi_C U_\alpha(t,s)=\prod_{l=1}^{m}\chi_{I_{\alpha l}}(p_l).$$

Proof: We are using the notation $F(p)u(x)=\widehat{F(p)\check{u}(p)}$ where $F:\mathbb{R}^n\to\mathbb{R}$, and $\hat{}$ and $\check{}$ are a Fourier transform and its inverse. The proof of Lemma 5 uses Proposition 1 and Lemma 4, i.e.,

$$\text{s-}\lim_{t\to\pm\infty}U_\alpha(s,t)\chi_C U_\alpha(t,s)$$
$$=\text{s-}\lim_{t\to\pm\infty}\exp(iX^t_s)\prod_{l=1}^{m}\exp[i(t-s)(-\Delta/2)]\,\chi_{B_{\alpha\beta}}\times\exp[-i(t-s)(-\Delta/2)]\exp(-iX^t_s)$$
$$=\text{s-}\lim_{t\to\pm\infty}\prod_{l=1}^{m}\exp(iX^t_s)\exp[-i|y_l|^2/2(t-s)]\,\chi_{B_{\alpha l}}\Big|_{y_l=p_l(t-s)}\times\exp[i|y_l|^2/2(t-s)]\exp(-iX^t_s)$$
$$=\prod_{l=1}^{m}\chi_{I_{\alpha l}}(p_l).$$

Lemma 6: If α and β are two channels with the same partition Γ, then

$$\text{s-}\lim_{t\to\pm\infty}U_\beta(s,t)\chi_C U_\alpha(t,s)-\exp[-i(t-s)(E_\alpha-E_\beta)]\prod_{l=1}^{m}\chi_{I_{\alpha l}}(p_l)=0.$$

Proof: Corollary to the proof of Lemma 5.

Lemma 7: Let $F(k,t):\mathbf{R}^{m+1}\to\mathbf{R}$ be such that $\lim_{t\to\pm\infty}(1/t^{|\xi|})D_k^\xi F(k,t)=0$ uniformly on compact sets for $1\leq|\xi|\leq m$, then $\lim_{t\to\pm\infty}\int_{\mathbf{R}^k}\exp[ikt+iF(k,t)]f(k)\,dk=0$ for any integrable function f.

Proof: It is sufficient to assume f is C^∞ with compact support. If $m=1$, then, by integration by parts,

$$\int_{\mathbf{R}}\exp(ikt)\exp[iF(k,t)]f(k)\,dk=\frac{\exp(ikt)}{it}\cdot\exp[iF(k,t)]\times f(k)\Big|_a^b-\int_a^b\frac{\exp(ikt)}{it}\,D_kF(k,t)\cdot\exp[iF(k,t)]f(k)+\exp[iF(k,t)]f'(k)dk-0$$

where support $f\subseteq[a,b]$. Just as integration by parts is based on $(hg)'=h'g+g'h$, for general m, use integration by parts based on

$$\frac{\partial^m}{\partial k_1\cdots\partial k_m}(HG)=\sum_{s=0}^{m}\sum_{\substack{\text{partition of}\\(k_1,\cdots,k_m)}}\frac{\partial^s H}{\partial k_{i_1}\cdots\partial k_{i_s}}\frac{\partial^{m-s}G}{\partial k_{i_{s+1}}\cdots\partial k_{i_m}}.$$

Solve for the term $s=m$ and integrate. Let $\partial^m H/\partial k_1\cdots\partial k_m=\exp(ikt)$ and $G=\exp[iF(k,t)]f(k)$. Note that $m-s$, the number of times $\exp(ikt)$ must be integrated in each of the other terms, is always greater than or equal to the number of derivatives of $F(k,t)$ that are present. (It could be greater because of the product rule on G.)

Lemma 8: If α and β are channels with distinct partitions, then

$$\text{w-}\lim_{t\to\pm\infty}U_\beta(s,t)\chi_c U_\alpha(t,s)=0.$$

Proof: Since $U_\beta(s,t)\chi_c U_\alpha(t,s)=U_\beta(s,t)U_\alpha(t,s)\times U_\alpha(s,t)\chi_c U(t,s)$, where $U_\alpha(s,t)\chi_c U_\alpha(t,s)$ converges strongly by Lemma 5, it need only be shown that $(U_\beta(s,t)U_\alpha(t,s)u,v)\to 0$. But this will follow from Lemma 7 and the assumption in Theorem 2. Compare Dollard[1] and Donaldson, Gibson, and Hersh.[9,10]

The evaluation of the series (3) now proceeds just as in Dollard's paper, and we do not reproduce it here. Briefly, Lemma 8 says the terms where α and β have distinct channels are zero by the Riemann–Lebesgue Lemma. Those terms where $\alpha\neq\beta$ have the same partition are zero by the orthogonal choice of bound states.[1] The terms where $\alpha=\beta$ are then evaluated by Lemma 5.

ACKNOWLEDGMENTS

I would like to mention the related work of Jauch, Lavine, and Newton[11] which generalizes Dollard's theorem by replacing $-\Delta=p^2$ by a more general function of p. I would also like to thank Dr. Jerome Goldstein and Dr. John Dollard for their very helpful comments and suggestions. This paper was written at the State University of New York, Binghamton, where the author was a Post-Doctoral Research Associate.

[1] J. D. Dollard, J. Math. Phys. **14**, 708 (1973).
[2] J. H. Hendrickson, J. Math. Phys. **4**, 768 (1975).
[3] P. Alsholm and T. Kato, "Proceedings of the Symposium on Pure and Applied Mathematics of The American Mathematical Society" (Am. Math. Soc., Providence, R.I., 1971), Vol. 23, pp. 393–99.
[4] J. A. Goldstein and C. J. Monlezun, "Temporally inhomogeneous scattering theory II: approximation theory and second order equations," SIAM J. Math. Anal., to appear.
[5] T. Kato, J. Math. Soc. Jpn. **5**, 208 (1953).
[6] J. D. Dollard, J. Math. Phys. **5**, 729 (1964).
[7] J. D. Dollard, Ph.D. dissertation, Princeton University.
[8] J. H. Hendrickson, Ph.D. dissertation, Tulane University (1974).
[9] P. Alshom, Ph.D. Dissertation, University of California, Berkeley.
[10] J. A. Donaldson, A. G. Gibson, and R. Hersh, J. Funct. Anal. **14**, 131-45 (1973).
[11] J. Jauch, R. Lavine, and R. Newton, Helv. Phys. Acta **45**, 325 (1972).

Stochastic wave-kinetic theory in the Liouville approximation*

Ioannis M. Besieris

Department of Electrical Engineering, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061

Frederick D. Tappert†

Courant Institute of Mathematical Sciences, 251 Mercer Street, New York, New York 10012
(Received 21 July 1975)

The behavior of scalar wave propagation in a wide class of asymptotically conservative, dispersive, weakly inhomogeneous and weakly nonstationary, anisotropic, random media is investigated on the basis of a stochastic, collisionless, Liouville-type equation governing the temporal evolution of a phase-space Wigner distribution density function. Within the framework of the first-order smoothing approximation, a general diffusion–convolution-type kinetic or transport equation is derived for the mean phase-space distribution function containing generalized (nonlocal, with memory) diffusion, friction, and absorption operators in phase space. Various levels of simplification are achieved by introducing additional constraints. In the long-time, Markovian, diffusion approximation, a general set of Fokker–Planck equations is derived. Finally, special cases of these equations are examined for spatially homogeneous systems and isotropic media.

I. INTRODUCTION

A general wave-kinetic method has been developed by Besieris and Tappert,[1–5] which makes possible the systematic derivation of generalized transport (or kinetic) equations that are valid even for partially coherent waves in inhomogeneous, dispersive, anisotropic media. The theory has been extended to include also media which are slowly varying in time and weakly absorbing (asymptotically conservative).

In order to examine wave propagation by means of this technique, a phase-space description of the problem is developed first. Using the concept of a general analytic signal for wave fields, a phase-space distribution function is defined following Wigner's phase-space approach to quantum mechanical waves. An exact equation of motion of this Wigner distribution function is derived next (it is referred to as the Wigner or wave-kinetic equation) which is fully equivalent to the original field equations. The concept of Weyl transforms (related to pseudo-differential operators) also plays an important role in the rigorous derivation of the phase-space description of wave propagation.

Randomness is introduced by considering the wave-kinetic equations for an ensemble of inhomogeneous media with specified statistical properties. Equations are derived for the temporal evolution of the ensemble-averaged phase-space distribution function from which physically meaningful average quantities are obtainable by taking appropriate phase-space moments.

The stochastic wave-kinetic method has already been used with success to study the behavior of scalar wave propagation in a wide class of random media, with applications to radar, sonar, and other types of communication systems. It generalizes the geometric optics approximation to include coherent effects such as diffraction and random and dispersive spreading of wave-packets. It also provides a systematic basis for many available classical transport or radiative transfer equations which have been formulated for the most part on the basis of ad hoc assumptions.

It is our intent in this paper to present a statistical analysis of the stochastic, collisionless Liouville equation

$$\frac{\partial}{\partial t} f(\mathbf{x},\mathbf{p},t;\alpha) = Lf(\mathbf{x},\mathbf{p},t;\alpha), \tag{1.1a}$$

$$Lf(\mathbf{x},\mathbf{p},t;\alpha) = \left(-\frac{\partial}{\partial \mathbf{p}}\,\omega_r(\mathbf{x},\mathbf{p},t;\alpha)\cdot\frac{\partial}{\partial \mathbf{x}} + \frac{\partial}{\partial \mathbf{x}}\,\omega_r(\mathbf{x},\mathbf{p},t;\alpha)\cdot\frac{\partial}{\partial \mathbf{p}} + 2\omega_i(\mathbf{x},\mathbf{p},t;\alpha)\right) f(\mathbf{x},\mathbf{p},t;\alpha). \tag{1.1b}$$

The medium is described by the linear dispersion relation $\omega(\mathbf{x},\mathbf{p},t;\alpha)$, and $\omega_r(\mathbf{x},\mathbf{p},t;\alpha)$ and $\omega_i(\mathbf{x},\mathbf{p},t;\alpha)$ are respectively the real and imaginary parts of $\omega(\mathbf{x},\mathbf{p},t;\alpha)$ (cf. Ref. 6). Within the framework of the wave-kinetic technique, $f(\mathbf{x},\mathbf{p},t;\alpha)$ is rigorously defined as the Weyl transform of the wave-analytic signal. Thus, $f(\mathbf{x},\mathbf{p},t;\alpha)$ is a Wigner distribution density function, and (1.1) follows from a systematic asymptotic expansion in the semiclassical (or correspondence-limit) approximation.

Equation (1.1) generalizes the Liouville-type stochastic partial differential equation arising in the geometric, ray-optical approach to random media.[7–11] Furthermore, when specialized to the one-species, linearized, Vlasov equation, it has been studied extensively[12] in connection with the stochastic acceleration of particles, a subject of importance in various areas such as cosmic rays, heating of thermonuclear plasma, turbulence of interstellar plasma, etc. Therefore, the results presented in this paper are also applicable to these fields.

In Sec. 2, a stochastic equation describing the evolution of the phase-space Wigner distribution function is derived and the conditions under which the Liouville approximation (1.1) is valid are discussed. In order for

the discussion to be self-contained, general equations for the mean and fluctuating parts of $f(\mathbf{x},\mathbf{p},t;\alpha)$ are derived in Sec. 3 using first a nonperturbative technique. These results are then specialized to the weak-coupling limit in order to arrive at the well-known first-order smoothing approximation. These findings are applied to the stochastic, collisionless Liouville equation (1.1) and a general diffusion-convolution-type kinetic or transport equation is given in Sec. 4. By using additional restrictions, various levels of approximation can be achieved. In Sec. 5, general Fokker–Planck equations for the mean phase-space distribution function are obtained. Finally, in Sec. 6, special cases of these general Fokker–Planck equations are considered for spatially homogeneous systems and isotropic media.

II. THE STOCHASTIC WAVE-KINETIC TECHNIQUE

A. The analytic signal

A large class of problems dealing with acoustic and electromagnetic scalar wave propagation in asymptotically conservative (cf. Lewis[13]), dispersive, weakly inhomogeneous and weakly nonstationary anisotropic media is governed by the general stochastic differential equation

$$\left\{\epsilon^2\left[\frac{\partial}{\partial t}-\Omega_i\left(\mathbf{x},t,-i\epsilon\frac{\partial}{\partial\mathbf{x}};\alpha\right)\right]^2+\Omega_r{}^2\left(\mathbf{x},t,-i\epsilon\frac{\partial}{\partial\mathbf{x}};\alpha\right)\right\}$$

$$\times u(\mathbf{x},t;\alpha)=0,\quad t\geq 0,\quad \mathbf{x}\in R^3. \tag{2.1}$$

A distinguishing feature of this problem is the presence of the positive dimensionless parameter ϵ, which can be taken to be inversely proportional to the scale size of the spatial and temporal irregularities. As such, for a slowly varying medium, ϵ will be a small but finite quantity. In (2.1), $\Omega_r{}^2$ is assumed to be a positive, self-adjoint, stochastic operator depending on a parameter $\alpha\in A$, A being a probability measure space. (All the fractional powers of $\Omega_r{}^2$ are defined and are, themselves, positive self-adjoint operators.) On the other hand, Ω_i which arises solely from the dissipative properties of the medium, is assumed to be a Hermitian stochastic operator. In addition, $u(\mathbf{x},t;\alpha)$, the real, scalar, random amplitude, is an element of an infinite-dimensional vector space. The problem is rendered closed by specifying Cauchy initial data and appropriate boundary conditions.

We shall be interested in the time evolution of observable quantities. In this sense, u and u^2 have little physical meaning. We may, however, consider the total wave energy and the total wave action which are given in terms of u, u_t and the operators Ω_r, Ω_i by integrals of the form

$$E=\tfrac{1}{2}\int_{R^3} F_1(u,u_t,\Omega_r,\Omega_i)\,d\mathbf{x}, \tag{2.2}$$

$$A=\tfrac{1}{2}\int_{R^3} F_2(u,u_t,\Omega_r,\Omega_i)\,d\mathbf{x}, \tag{2.3}$$

respectively. (In the absence of dissipation, $F_1=u\Omega_r{}^2u+\epsilon^2u_t{}^2$ and $F_2=u\Omega_ru+\epsilon^2u_t\Omega_r^{-1}u_t$). In view of the assumption that the medium is time-dependent and dissipative, neither of these quantities is conserved. The integrands in (2.2) and (2.3) are, respectively, the space wave energy and wave action density functions.

The difficulty of working with the complicated expressions (2.2) and (2.3) directly is circumvented by introducing the notion of the *complex analytic signal*. This quantity is defined by means of the relation

$$\psi(\mathbf{x},t;\alpha)=2^{-1/2}(\Omega_r^{1/2}u+i\epsilon\Omega_r^{-1/2}D_tu), \tag{2.4}$$

with the operator D_t given by $D_t=(\partial/\partial t)-\Omega_i[\mathbf{x},t,-i\epsilon(\partial/\partial\mathbf{x});\alpha]$. The total wave energy and wave action associated with (2.1) are given in terms of the analytic signal as follows:

$$E=\int_{R^3}\psi^*\Omega_r\psi\,d\mathbf{x}, \tag{2.5}$$

$$A=\int_{R^3}\psi^*\psi\,d\mathbf{x}. \tag{2.6}$$

Taking the time derivative of (2.4) and using (2.1), one has the formal relation

$$i\epsilon D_t\psi=\Omega_r\psi+\tfrac{1}{4}i\epsilon D_t(\log\Omega_r{}^{1/2})\psi^*. \tag{2.7}$$

By neglecting nonadiabatic terms, the analytic signal obeys the closed equation

$$i\epsilon\frac{\partial}{\partial t}\psi=\Omega_r\psi+i\epsilon\,\Omega_i\psi. \tag{2.8}$$

(In the absence of dissipation, the total wave action is an adiabatic invariant to all orders in ϵ; E, however, is not conserved because of the time dependence of the medium.)

B. The Wigner distribution function

The two-point, equal time density function is introduced next as follows in terms of the analytic signal:

$$\rho(\mathbf{x}_2,\mathbf{x}_1,t;\alpha)=\psi^*(\mathbf{x}_2,t;\alpha)\psi(\mathbf{x}_1,t;\alpha). \tag{2.9}$$

It obeys the von Neumann-like equation

$$i\epsilon\frac{\partial}{\partial t}\rho=-[\Omega_r,\rho]_-+i\epsilon[\Omega_i,\rho]_+, \tag{2.10}$$

where $[A,B]_\mp=AB\mp BA$ denote the usual commutation and anticommutation relations.

The phase-space analog of the density function is provided by the Wigner distribution function which is defined as follows[14]:

$$f(\mathbf{x},\mathbf{p},t;\alpha)=(2\pi\epsilon)^{-3}\int_{R^3}d\mathbf{y}\exp(i\mathbf{p}\cdot\mathbf{y}/\epsilon)\rho(\mathbf{x}+\tfrac{1}{2}\mathbf{y},\mathbf{x}-\tfrac{1}{2}\mathbf{y},t;\alpha). \tag{2.11}$$

This quantity is real, but not necessarily positive everywhere. In this sense, it does not qualify as a *bona fide* probability density function. One may introduce it in terms of other bilinear expressions of the analytic signal. The relation (2.11) has been chosen because of its relative simplicity and symmetry. The total wave energy and wave action can be written in terms of the Wigner distribution function as follows:

$$E=\int_{R^3}d\mathbf{x}\int_{R^3}d\mathbf{p}\;\omega_r(\mathbf{x},\mathbf{p},t;\alpha)f(\mathbf{x},\mathbf{p},t;\alpha), \tag{2.12}$$

$$A=\int_{R^3}d\mathbf{x}\int_{R^3}d\mathbf{p}f(\mathbf{x},\mathbf{p},t;\alpha). \tag{2.13}$$

Here, $\omega_r(\mathbf{x},\mathbf{p},t;\alpha)$ is the Weyl transform of the operator Ω_r. By virtue of (2.12), $\omega_r f$ can be interpreted as the wave energy density in phase space. Similarly, from (2.13), f can be thought of as the wave action density in phase space.

Given the definition of $f(\mathbf{x},\mathbf{p},t;\alpha)$ and using the von Neumann equation (2.10), it is found that the Wigner distribution function evolves according to the following equation[15]:

$$\frac{\partial}{\partial t}f(\mathbf{x},\mathbf{p},t;\alpha) = \omega_r(\mathbf{x},\mathbf{p},t;\alpha)\frac{2}{\epsilon}\sin\left[\frac{\epsilon}{2}\left(\frac{\overleftarrow{\partial}}{\partial\mathbf{x}}\cdot\frac{\overrightarrow{\partial}}{\partial\mathbf{p}}-\frac{\overleftarrow{\partial}}{\partial\mathbf{p}}\cdot\frac{\overrightarrow{\partial}}{\partial\mathbf{x}}\right)\right]f(\mathbf{x},\mathbf{p},t;\alpha) + \omega_i(\mathbf{x},\mathbf{p},t;\alpha)\,2\cos\left[\frac{\epsilon}{2}\left(\frac{\overleftarrow{\partial}}{\partial\mathbf{x}}\cdot\frac{\overrightarrow{\partial}}{\partial\mathbf{p}}-\frac{\overleftarrow{\partial}}{\partial\mathbf{p}}\cdot\frac{\overrightarrow{\partial}}{\partial\mathbf{x}}\right)\right]f(\mathbf{x},\mathbf{p},t;\alpha), \tag{2.14}$$

where $\omega_i(\mathbf{x},\mathbf{p},t;\alpha)$ is the Weyl transform of the operator Ω_i. (Depending on the directions of the arrows, the differential operators on the right-hand side of (2.14) operate on ω_r, ω_i, or f.) We shall refer to the exact equation of evolution of f as the *stochastic Wigner equation*.

It is seen from (2.14) that in the "correspondence limit" ($\epsilon \to 0$), the Wigner distribution function obeys the simpler relation

$$\frac{\partial}{\partial t}f(\mathbf{x},\mathbf{p},t;\alpha) = Lf(\mathbf{x},\mathbf{p},t;\alpha), \tag{2.15a}$$

$$Lf(\mathbf{x},\mathbf{p},t;\alpha) = \left(-\frac{\partial}{\partial\mathbf{p}}\,\omega_r(\mathbf{x},\mathbf{p},t;\alpha)\cdot\frac{\partial}{\partial\mathbf{x}}+\frac{\partial}{\partial\mathbf{x}}\,\omega_r(\mathbf{x},\mathbf{p},t;\alpha)\cdot\frac{\partial}{\partial\mathbf{p}}+2\omega_i(\mathbf{x},\mathbf{p},t;\alpha)\right)f(\mathbf{x},\mathbf{p},t;\alpha)+O(\epsilon^2). \tag{2.15b}$$

In the framework of this approximation, we shall refer to (2.15) as the *stochastic, collisionless Liouville equation*.

III. GENERAL EQUATIONS FOR THE MEAN AND FLUCTUATING FIELDS

In the first part of this section we derive general equations for the mean and fluctuating parts of the Wigner distribution function using a nonperturbative statistical approach. These results are then specialized to the weak-coupling limit in order to arrive at the well known first-order smoothing approximation. Further simplifications, required for the subsequent development, lead to the long-time and Markovian approximations. The discussion in this section is general and applies to both the stochastic Wigner equation [cf. Eq. (2.14)] and the stochastic, collisionless Liouville equation [cf. Eq. (2.15)].

Consider the linear, stochastic, partial different equation[16]

$$\frac{\partial}{\partial t}f(\mathbf{x},\mathbf{p},t;\alpha) = Lf(\mathbf{x},\mathbf{p},t;\alpha), \tag{3.1a}$$

$$f(\mathbf{x},\mathbf{p},0;\alpha) = f_0(\mathbf{x},\mathbf{p};\alpha). \tag{3.1b}$$

The stochastic operator L is split into two parts as folows: $L = L_0 + L_1$. The selection of L_0 and L_1 is made in such a way that they are linear operators in an infinite dimensional vector space H, corresponding, respectively, to "free" and "interaction" propagation.

The distribution function f is, in turn, decomposed abstractly into two mutually independent terms, viz., $f = Vf + Cf$ by means of the formal introduction of the two operators V and C.[17] Vf is called the *mean* component, and Cf is the *fluctuating* part of the distribution function f.[18] The uniqueness of the decomposition as well as the mutual independence of the two components are ensured by prescribing the properties $V + C = I$, $V^2 = V$, $C^2 = C$, $VC = 0$, $CV = 0$, where I is the identity operator. By virtue of these relations, V and C are called *projection operators*.

The interconnection between the decompositions for the operator L and the distribution function f is contained in the commutation relations $[L_0, V]_- = 0$ and $[L_0, C]_- = 0$ which constitute a mathematical statement of the fact that the fluctuating part of f is due only to the interaction part of the operator L. Therefore, L_0 must commute with V, and also, with $C = I - V$.

The specific realization of the projection operators V and C which will be used in the ensuing work is the following: $Vf \to E\{f\}$, $Cf \to \delta f$, where $E\{f\}$ and δf are, respectively, the ensemble average and fluctuating part of the random distribution $f(\alpha)$. Within the framework of this specific realization, the aforementioned commutation relations signify that L_0 is a deterministic operator and L_1 is a generally noncentered random operator.

A. Equations for the mean and fluctuating distribution functions; first-order smoothing approximation

Operating on (3.1) with the projection operators V, C yields the equations

$$\frac{\partial}{\partial t}E\{f(t)\} = VLE\{f(t)\} + VL\,\delta f(t), \tag{3.2a}$$

$$\frac{\partial}{\partial t}\delta f(t) = CL\,\delta f(t) + CLE\{f(t)\}, \tag{3.2b}$$

respectively. Equation (3.2b) can be solved for $\delta f(t)$ in terms of the mean field and the initial value of the fluctuating part of the distribution:

$$\delta f(t) = U_V(t,0)\,\delta f(0) + \int_0^t dt_1 U_V(t,t_1)CL(t_1)E\{f(t_1)\}. \tag{3.3}$$

The propagator U_V is defined as the solution of the initial value problem

$$\frac{\partial}{\partial t}U_V(t,0) = CL(t)U_V(t,0), \quad U_V(0,0) = 1. \tag{3.4}$$

Substituting (3.3) into (3.2a) results in the equation[19]

$$\frac{\partial}{\partial t}E\{f(t)\} = L_0E\{f(t)\} + VL_1VE\{f(t)\} + VL_1U_V(t,0)C\,\delta f(0) + \int_0^t dt_1 VL_1(t)U_V(t,t_1)CL_1(t_1)E\{f(t_1)\}. \tag{3.5}$$

This formal expression for the mean field is valid for both weak and strong fluctuations in the randomly vary-

ing inhomogeneities of the medium. It should also be noted that no restriction whatsoever has been imposed on the random operator L_1 and the initial value of the distribution function.

Equation (3.3) indicates that the fluctuating part of the field can be calculated by quadratures once the mean distribution function has been determined separately.

B. The first-order smoothing approximation

Balescu and Misguich[20] have shown recently that

$$U_V(t,0)=\sum_{n=0}^{\infty}\left[\int_0^t dt_1 W(t,t_1)CL_1(t_1)\right]^n W(t,0), \tag{3.6}$$

with the propagator W defined as the solution of the initial value problem

$$\frac{\partial}{\partial t}W(t,0)=L_0(t)W(t,0),\quad W(0,0)=1. \tag{3.7}$$

The *first-order smoothing approximation* is determined by introducing in (3.5) the weak-coupling limit approximation

$$U_V(t,0)C\underset{n=0}{\rightarrow}CW(t,0). \tag{3.8}$$

For the sake of simplicity, we also impose the restriction that L_1 is a centered random operator. This condition is stated mathematically as $VL_1V=0$. Furthermore, $f(0)$ is taken to be deterministic so that $C\delta f(0)=0$. We have, then, in the place of (3.5)

$$\frac{\partial}{\partial t}E\{f(t)\} \tag{3.9}$$
$$=L_0E\{f(t)\}+\int_0^t dt_1 VL_1(t)W(t,t_1)L_1(t_1)E\{f(t_1)\}.$$

The first-order smoothing approximation (cf. also Refs. 9, 10, 21–23) is essentially a perturbational method applicable for weak random variations. The mean field $E\{f(t)\}$, which is determined by successive iterations of (3.9), is found to be a partial summation of the exact, infinite, conventional perturbation series solution. This subseries, besides yielding results consistent with physically imposed constraints, enables one to circumvent certain time and space secularities which are characteristic of solutions consisting of only a finite number of terms of the infinite perturbation series (e.g., Born approximation and its various modifications).

In general, the solution of (3.7) for the propagator W needed in (3.9) has the form

$$W(t,0)=X\exp\left\{\int_0^t dt_1 L_0(t_1)\right\}, \tag{3.10}$$

where X is a time-ordering operator. In the following, we shall assume that L_0 is time-independent. In this case (3.10) simplifies to

$$W(t,0)=\exp(tL_0). \tag{3.11}$$

The approximation (3.9) together with the assumption (3.11) will be used in the following section to derive a diffusion-convolution-type kinetic equation for the average part of the Wigner distribution function.

C. Long-time Markovian approximation

Let $E\{\hat{f}(t)\}$ denote the asymptotic limit of the field $E\{f(t)\}$ as $t\to\infty$. Then, given that $U_V(t,0)C\to 0$ as $t\to\infty$, it can be established that (3.5) assumes the simpler form[17,24]

$$\frac{\partial}{\partial t}E\{\hat{f}(t)\} \tag{3.12}$$
$$=L_0E\{\hat{f}(t)\}+\int_0^\infty dt_1 VL_1(t)U_V(t,t_1)CL_1(t_1)E\{\hat{f}(t_1)\}.$$

We shall call this relation the *long-time approximation* corresponding to (3.5). It is interesting to note that the term in (3.5) containing the fluctuating part of the initial distribution is asymptotically null in this approximation. No restriction need, therefore, be imposed on $f(0)$. It should, further, be pointed out that the initial mean distribution $E\{\hat{f}(0)\}$ required for the complete specification of the initial value problem (3.12) must be chosen so that the solution of (3.12) will coincide, for large times, with the asymptotic value of the solution of the "exact" equation (3.5) with the initial value $E\{f(0)\}$.

In the first-order smoothing approximation and under the assumption that the background deterministic medium is time-independent, (3.12) simplifies to

$$\frac{\partial}{\partial t}E\{\hat{f}(t)\}=L_0E\{\hat{f}(t)\} \tag{3.13}$$
$$+\int_0^\infty dt_1 VL_1(t)W(t_1)L_1(t-t_1)E\{\hat{f}(t-t_1)\}.$$

This relation can be rewritten in the equivalent, purely differential form

$$\frac{\partial}{\partial t}E\{\hat{f}(t)\}=KE\{\hat{f}(t)\}, \tag{3.14}$$

K being the solution of the nonlinear integral equation

$$K=L_0+\int_0^\infty dt_1 VL_1(t)W(t_1)L_1(t-t_1)\exp(-t_1K). \tag{3.15}$$

The last expression can be solved for K by the method of successive substitutions. If only the first two terms of the expansion are retained, one obtains the *long-time Markovian approximation*

$$\frac{\partial}{\partial t}E\{\hat{f}(t)\}=L_0E\{\hat{f}(t)\}+\int_0^\infty dt_1 VL_1(t)W(t_1)L_1(t-t_1)E\{\hat{f}(t)\}. \tag{3.16}$$

This simplified integro-differential equation for the mean field will be used in Sec. 5 to derive a general Fokker–Planck equation.

IV. KINETIC EQUATION FOR THE MEAN DISTRIBUTION FUNCTION

In this section we specialize our findings in the previous section to the case of the stochastic Liouville-type operator

$$L=\frac{\partial}{\partial \mathbf{x}}\omega_r(\mathbf{x},\mathbf{p};\alpha)\cdot\frac{\partial}{\partial \mathbf{p}}-\frac{\partial}{\partial \mathbf{p}}\omega_r(\mathbf{x},\mathbf{p};\alpha)\cdot\frac{\partial}{\partial \mathbf{x}}+2\omega_i(\mathbf{x},\mathbf{p};\alpha) \tag{4.1}$$

introduced earlier [cf. (2.15)] in connection with the stochastic wave-kinetic technique. For simplicity, L is assumed in the sequel to be independent of time. This restriction is not a serious drawback since it can be

easily lifted (cf. Ref. 5). Thus, L is translationally invariant with respect to time, and its free and interaction parts are given simply by

$$L_0 = \frac{\partial}{\partial \mathbf{x}} E\{\omega_r(\mathbf{x}, \mathrm{p}; \alpha)\} \cdot \frac{\partial}{\partial \mathrm{p}} - \frac{\partial}{\partial \mathrm{p}} E\{\omega_r(\mathbf{x}, \mathrm{p}; \alpha)\} \circ \frac{\partial}{\partial \mathbf{x}} + 2E\{\omega_i(\mathbf{x}, \mathrm{p}; \alpha)\}, \tag{4.2a}$$

$$L_1 = \frac{\partial}{\partial \mathbf{x}} \delta\omega_r(\mathbf{x}, \mathrm{p}; \alpha) \circ \frac{\partial}{\partial \mathrm{p}} - \frac{\partial}{\partial \mathrm{p}} \delta\omega_r(\mathbf{x}, \mathrm{p}; \alpha) \cdot \frac{\partial}{\partial \mathbf{x}} + 2\delta\omega_i(\mathbf{x}, \mathrm{p}; \alpha) \tag{4.2b}$$

in terms of the mean and the fluctuating parts of ω_r and ω_i.

Diffusion-convolution-type kinetic equation

We commence with the first-order smoothing approximation of the Dyson-Schwinger equation for the mean Wigner distribution function [cf. Eq. (3.9)]. In order to write our explicitly the second part on the right-hand side of (3.9), we use the definition of L_1 given in (4.2b). We have, then

$$\begin{aligned}
\left(\frac{\partial}{\partial t} - L_0\right) E\{f(\mathbf{x}, \mathrm{p}, t; \alpha)\} = {} & \frac{\partial}{\partial \mathrm{p}} \circ \left(\int_0^t dt_1 E\left\{ \frac{\partial}{\partial \mathbf{x}} \delta\omega_r(\mathbf{x}, \mathrm{p}; \alpha) \exp(t_1 L_0) \frac{\partial}{\partial \mathbf{x}} \delta\omega_r(\mathbf{x}, \mathrm{p}; \alpha) \right\} \circ \frac{\partial}{\partial \mathrm{p}} E\{f(\mathbf{x}, \mathrm{p}, t - t_1; \alpha)\} \right) \\
& - \frac{\partial}{\partial \mathrm{p}} \circ \left(\int_0^t dt_1 E\left\{ \frac{\partial}{\partial \mathbf{x}} \delta\omega_r(\mathbf{x}, \mathrm{p}; \alpha) \exp(t_1 L_0) \frac{\partial}{\partial \mathrm{p}} \delta\omega_r(\mathbf{x}, \mathrm{p}; \alpha) \right\} \circ \frac{\partial}{\partial \mathbf{x}} E\{f(\mathbf{x}, \mathrm{p}, t - t_1; \alpha)\} \right) \\
& - \frac{\partial}{\partial \mathbf{x}} \circ \left(\int_0^t dt_1 E\left\{ \frac{\partial}{\partial \mathrm{p}} \delta\omega_r(\mathbf{x}, \mathrm{p}; \alpha) \exp(t_1 L_0) \frac{\partial}{\partial \mathbf{x}} \delta\omega_r(\mathbf{x}, \mathrm{p}; \alpha) \right\} \circ \frac{\partial}{\partial \mathrm{p}} E\{f(\mathbf{x}, \mathrm{p}, t - t_1; \alpha)\} \right) \\
& + \frac{\partial}{\partial \mathbf{x}} \circ \left(\int_0^t dt_1 E\left\{ \frac{\partial}{\partial \mathrm{p}} \delta\omega_r(\mathbf{x}, \mathrm{p}; \alpha) \exp(t_1 L_0) \frac{\partial}{\partial \mathrm{p}} \delta\omega_r(\mathbf{x}, \mathrm{p}; \alpha) \right\} \circ \frac{\partial}{\partial \mathbf{x}} E\{f(\mathbf{x}, \mathrm{p}, t - t_1; \alpha)\} \right) \\
& + \frac{\partial}{\partial \mathrm{p}} \cdot \left(\int_0^t dt_1 E\left\{ \frac{\partial}{\partial \mathbf{x}} \delta\omega_r(\mathbf{x}, \mathrm{p}; \alpha) \exp(t_1 L_0) 2\delta\omega_i(\mathbf{x}, \mathrm{p}; \alpha) \right\} E\{f(\mathbf{x}, \mathrm{p}, t - t_1; \alpha)\} \right) \\
& - \frac{\partial}{\partial \mathbf{x}} \cdot \left(\int_0^t dt_1 E\left\{ \frac{\partial}{\partial \mathrm{p}} \delta\omega_r(\mathbf{x}, \mathrm{p}; \alpha) \exp(t_1 L_0) 2\delta\omega_i(\mathbf{x}, \mathrm{p}; \alpha) \right\} E\{f(\mathbf{x}, \mathrm{p}, t - t_1; \alpha)\} \right) \\
& + \int_0^t dt_1 E\left\{ 2\delta\omega_i(\mathbf{x}, \mathrm{p}; \alpha) \exp(t_1 L_0) \frac{\partial}{\partial \mathbf{x}} \delta\omega_r(\mathbf{x}, \mathrm{p}; \alpha) \right\} \cdot \frac{\partial}{\partial \mathrm{p}} E\{f(\mathbf{x}, \mathrm{p}, t - t_1; \alpha)\} \\
& - \int_0^t dt_1 E\left\{ 2\delta\omega_i(\mathbf{x}, \mathrm{p}; \alpha) \exp(t_1 L_0) \frac{\partial}{\partial \mathrm{p}} \delta\omega_r(\mathbf{x}, \mathrm{p}; \alpha) \right\} \cdot \frac{\partial}{\partial \mathbf{x}} E\{f(\mathbf{x}, \mathrm{p}, t - t_1; \alpha)\} \\
& + \int_0^t dt_1 E\{2\delta\omega_i(\mathbf{x}, \mathrm{p}; \alpha) \exp(t_1 L_0) 2\,\delta\omega_i(\mathbf{x}, \mathrm{p}; \alpha)\} E\{f(\mathbf{x}, \mathrm{p}, t - t_1; \alpha)\}.
\end{aligned} \tag{4.3}$$

This rather formidable integro-differential relation, which will be referred to as the *kinetic equation* for the mean distribution function, generalizes previous equations of this type (cf. Refs. 7–12). It applies to media with inhomogeneous deterministic background, and constitutes a uniform approximation, valid for any value of time, from which short and long time limiting cases can be considered. (The latter will be dealt with in detail in the following section). The right-hand side of (4.3) contains generalized diffusion operators (nonlocal, with memory) in phase space, and, also, generalized friction and absorption operators.

For a spatially homogeneous background, (4.3) is a generalization of the kinetic equation for random geometrical optics obtained by Frisch (cf. Ref. 10). It is a *diffusion-type equation* with respect to space and wave vector (momentum) coordinates, and a *convolution equation* with respect to position and time. (Under various special conditions, e.g., homogeneous and isotropic randomness and spatially homogeneous systems, it may be possible to arrive at an exact analytical solution.)

V. GENERAL FOKKER-PLANCK EQUATION

A. Diffusion, friction, and absorption coefficients

By imposing additional restrictions, various levels of simplification of (4.3) can be obtained. The long-time, Markovian approximation yields the expression

$$\begin{aligned}
\left(\frac{\partial}{\partial t} + \mathrm{v}(\mathrm{p}) \cdot \frac{\partial}{\partial \mathbf{x}}\right) E\{\hat{f}(\mathbf{x}, \mathrm{p}, t; \alpha)\} = {} & \frac{\partial}{\partial \mathrm{p}} \cdot \left[\mathbf{D}^{rr}_{\mathrm{pp}} \cdot \frac{\partial}{\partial \mathrm{p}} E\{\hat{f}(\mathbf{x}, \mathrm{p}, t; \alpha)\} \right] \\
& - \frac{\partial}{\partial \mathrm{p}} \cdot \left[\mathbf{D}^{rr}_{\mathrm{px}} \cdot \frac{\partial}{\partial \mathbf{x}} E\{\hat{f}(\mathbf{x}, \mathrm{p}, t; \alpha)\} \right] \\
& - \frac{\partial}{\partial \mathbf{x}} \cdot \left[\mathbf{D}^{rr}_{\mathrm{xp}} \cdot \frac{\partial}{\partial \mathrm{p}} E\{\hat{f}(\mathbf{x}, \mathrm{p}, t; \alpha)\} \right] \\
& + \frac{\partial}{\partial \mathbf{x}} \circ \left[\mathbf{D}^{rr}_{\mathrm{xx}} \cdot \frac{\partial}{\partial \mathbf{x}} E\{\hat{f}(\mathbf{x}, \mathrm{p}, t; \alpha)\} \right] \\
& + \frac{\partial}{\partial \mathrm{p}} \circ [\mathbf{F}^{ri}_{p} E\{\hat{f}(\mathbf{x}, \mathrm{p}, t; \alpha)\}] \\
& - \frac{\partial}{\partial \mathbf{x}} \circ [\mathbf{F}^{ri}_{x} E\{\hat{f}(\mathbf{x}, \mathrm{p}, t; \alpha)\}] \\
& + \mathbf{F}^{ir}_{p} \circ \frac{\partial}{\partial \mathrm{p}} E\{\hat{f}(\mathbf{x}, \mathrm{p}, t; \alpha)\} - \mathbf{F}^{ir}_{x} \circ \frac{\partial}{\partial \mathbf{x}} E\{\hat{f}(\mathbf{x}, \mathrm{p}, t; \alpha)\} \\
& + A^{ii} E\{\hat{f}(\mathbf{x}, \mathrm{p}, t; \alpha)\} + A^{i} E\{\hat{f}(\mathbf{x}, \mathrm{p}, t; \alpha)\},
\end{aligned} \tag{5.1}$$

with the dyadic (**D**), vector (**F**), and scalar operator coefficients defined by

$$\mathbf{D}_{pp}^{rr}=\int_0^\infty d\tau\, E\left\{\frac{\partial}{\partial \mathbf{x}}\delta\omega_r(\mathbf{x},\mathbf{p};\alpha)\times\exp\left(-\tau\mathbf{v}\cdot\frac{\partial}{\partial\mathbf{x}}\right)\frac{\partial}{\partial\mathbf{x}}\delta\omega_r(\mathbf{x},\mathbf{p};\alpha)\right\} \tag{5.2a}$$

$$\mathbf{D}_{px}^{rr}=\int_0^\infty d\tau\, E\left\{\frac{\partial}{\partial \mathbf{x}}\delta\omega_r(\mathbf{x},\mathbf{p};\alpha)\times\exp\left(-\tau\mathbf{v}\cdot\frac{\partial}{\partial\mathbf{x}}\right)\frac{\partial}{\partial\mathbf{p}}\delta\omega_r(\mathbf{x},\mathbf{p};\alpha)\right\}, \tag{5.2b}$$

$$\mathbf{D}_{xp}^{rr}=\int_0^\infty d\tau\, E\left\{\frac{\partial}{\partial \mathbf{p}}\delta\omega_r(\mathbf{x},\mathbf{p};\alpha)\times\exp\left(-\tau\mathbf{v}\cdot\frac{\partial}{\partial\mathbf{x}}\right)\frac{\partial}{\partial\mathbf{x}}\delta\omega_r(\mathbf{x},\mathbf{p};\alpha)\right\}, \tag{5.2c}$$

$$\mathbf{D}_{xx}^{rr}=\int_0^\infty d\tau\, E\left\{\frac{\partial}{\partial \mathbf{p}}\delta\omega_r(\mathbf{x},\mathbf{p};\alpha)\times\exp\left(-\tau\mathbf{v}\cdot\frac{\partial}{\partial\mathbf{x}}\right)\frac{\partial}{\partial\mathbf{p}}\delta\omega_r(\mathbf{x},\mathbf{p};\alpha)\right\}, \tag{5.2d}$$

$$\mathbf{F}_{p}^{ri}=\int_0^\infty d\tau\, E\left\{\frac{\partial}{\partial \mathbf{x}}\delta\omega_r(\mathbf{x},\mathbf{p};\alpha)\times\exp\left(-\tau\mathbf{v}\cdot\frac{\partial}{\partial\mathbf{x}}\right)2\delta\omega_i(\mathbf{x},\mathbf{p};\alpha)\right\}, \tag{5.2e}$$

$$\mathbf{F}_{x}^{ri}=\int_0^\infty d\tau\, E\left\{\frac{\partial}{\partial \mathbf{p}}\delta\omega_r(\mathbf{x},\mathbf{p};\alpha)\times\exp\left(-\tau\mathbf{v}\cdot\frac{\partial}{\partial\mathbf{x}}\right)2\delta\omega_i(\mathbf{x},\mathbf{p};\alpha)\right\}, \tag{5.2f}$$

$$\mathbf{F}_{p}^{ir}=\int_0^\infty d\tau\, E\left\{2\delta\omega_i(\mathbf{x},\mathbf{p};\alpha)\times\exp\left(-\tau\mathbf{v}\cdot\frac{\partial}{\partial\mathbf{x}}\right)\frac{\partial}{\partial\mathbf{x}}\delta\omega_r(\mathbf{x},\mathbf{p};\alpha)\right\}, \tag{5.2g}$$

$$\mathbf{F}_{x}^{ir}=\int_0^\infty d\tau\, E\left\{2\delta\omega_i(\mathbf{x},\mathbf{p};\alpha)\times\exp\left(-\tau\mathbf{v}\cdot\frac{\partial}{\partial\mathbf{x}}\right)\frac{\partial}{\partial\mathbf{p}}\delta\omega_r(\mathbf{x},\mathbf{p};\alpha)\right\}, \tag{5.2h}$$

$$A^{ii}=\int_0^\infty d\tau\, E\left\{2\delta\omega_i(\mathbf{x},\mathbf{p};\alpha)\times\exp\left(-\tau\mathbf{v}\cdot\frac{\partial}{\partial\mathbf{x}}\right)2\delta\omega_i(\mathbf{x},\mathbf{p};\alpha)\right\}, \tag{5.2i}$$

$$A^{i}=2\,E\{\omega_i(\mathbf{x},\mathbf{p};\alpha)\} \tag{5.2j}$$

and the vector quantity $\mathbf{v}(\mathbf{p})$ given as follows:

$$\mathbf{v}(\mathbf{p})=\frac{\partial}{\partial\mathbf{p}}E\{\omega_r(\mathbf{x},\mathbf{p};\alpha)\}. \tag{5.3}$$

In writing down (5.1) we have resorted to the following simplifying assumptions:

(i) The deterministic background medium is spatially homogeneous. (This implies that both $E\{\omega_r(\mathbf{x},\mathbf{p};\alpha)\}$ and $E\{\omega_i(\mathbf{x},\mathbf{p};\alpha)\}$ are independent of position.)

(ii) The quantity $E\{\omega_i(\mathbf{x},\mathbf{p};\alpha)\}$ is a slowly varying function of momentum so that its first- and higher-order derivatives with respect to $\mathbf{p}$ can be neglected by comparison to the quantity itself.

In the following, the translational effects of the operator $\exp[-\tau\mathbf{v}\cdot(\partial/\partial\mathbf{x})]$ on the mean distribution function $E\{f(\mathbf{x},\mathbf{p},t;\alpha)\}$ will be ignored. This simplication is known as the *diffusion approximation*. Within the confines of this approximation, (5.1) becomes a Fokker–Planck equation. The coefficients **D** are called the *dyadic diffusion coefficients*, the quantities **F** are the *vector friction coefficients*, and, finally, the A's are the *scalar absorption coefficients*.

B. Anisotropic, dissipative and/or dispersive systems; uniform, homogeneous, and isotropic fluctuations

We shall derive here explicit expressions for the diffusion, friction, and absorption coefficients (5.2) in the case of a general anisotropic, dissipative and/or dispersive medium characterized by uniform, spatially homogeneous and isotropic fluctuations of the randomly varying functions $\delta\omega_r(\mathbf{x},\mathbf{p};\alpha)$ and $\delta\omega_i(\mathbf{x},\mathbf{p};\alpha)$. We assume, furthermore, the conditions (i) and (ii) specified in the previous subsection. Thus, the background medium is independent of the space coordinates, and $E\{\omega_i(\mathbf{x},\mathbf{p};\alpha)\}$ is a slowly varying function of $\mathbf{p}$.

In many physically important problems $\delta\omega_r(\mathbf{x},\mathbf{p};\alpha)$ can be written as a product of a random function $\delta\chi_r(\mathbf{x};\alpha)$, and a deterministic function $g_r(\mathbf{p})$, viz.,

$$\delta\omega_r(\mathbf{x},\mathbf{p};\alpha)=\delta\chi_r(\mathbf{x};\alpha)g_r(\mathbf{p}). \tag{5.4}$$

Similarly, $\delta\omega_i(\mathbf{x},\mathbf{p};\alpha)$ is written as a product of a random function of position, $\delta\chi_i(\mathbf{x};\alpha)$, and a deterministic function, $g_i(\mathbf{p})$, viz.,

$$\delta\omega_i(\mathbf{x},\mathbf{p};\alpha)=\delta\chi_i(\mathbf{x};\alpha)g_i(\mathbf{p}). \tag{5.5}$$

For spatially homogeneous and isotropic fluctuations, we define the following two-point correlation functions:

$$\Gamma_{rr}(y)=E\{\delta\chi_r(\mathbf{x};\alpha)\delta\chi_r(\mathbf{x}-\mathbf{y};\alpha)\}, \tag{5.6a}$$

$$\Gamma_{ri}(y)=E\{\delta\chi_r(\mathbf{x};\alpha)\delta\chi_i(\mathbf{x}-\mathbf{y};\alpha)\}, \tag{5.6b}$$

$$\Gamma_{ir}(y)=E\{\delta\chi_i(\mathbf{x};\alpha)\delta\chi_r(\mathbf{x}-\mathbf{y};\alpha)\}, \tag{5.6c}$$

$$\Gamma_{ii}(y)=E\{\delta\chi_i(\mathbf{x};\alpha)\delta\chi_i(\mathbf{x}-\mathbf{y};\alpha)\}. \tag{5.6d}$$

Furthermore, for a uniform, spatially homogeneous and isotropic model, we specify the relationships:

$$\frac{\Gamma_{rr}(y)}{\Gamma_{rr}(0)}=\frac{\Gamma_{ri}(y)}{\Gamma_{ri}(0)}=\frac{\Gamma_{ir}(y)}{\Gamma_{ir}(0)}=\frac{\Gamma_{ii}(y)}{\Gamma_{ii}(0)}=\rho(y), \tag{5.7a}$$

$$\rho(0)=1. \tag{5.7b}$$

The quantity $\rho(y)$ will be referred to as the correlation coefficient.

While, admittedly, more complicated choices for the correlation functions may more closely resemble the true situations in the real world, the relatively simple forms (5.6) and (5.7) contain the essential behavior of the random fluctuations.

With $y=v\tau$, where $v=|\mathbf{v}|$, the first dyadic diffusion coefficient becomes

$$\mathbf{D}_{pp}^{rr}=-\frac{1}{v(\mathbf{p})}g_r^2(\mathbf{p})\Gamma_{rr}(0)\int_0^\infty dy\left[\left(\mathbf{I}-\frac{\mathbf{vv}}{v^2}\right)\frac{1}{y}\frac{\partial}{\partial y}\rho(y)-\frac{\mathbf{vv}}{v^2}\frac{\partial^2}{\partial y^2}\rho(y)\right], \tag{5.8}$$

which, in turn, simplifies to

$$\mathbf{D}_{pp}^{rr}=\frac{1}{v(\mathbf{p})}g_r^2(\mathbf{p})\left(\mathbf{I}-\frac{\mathbf{vv}}{v^2}\right)\Gamma_{rr}(0)B, \tag{5.9a}$$

$$B=-\int_0^\infty dy\,\frac{1}{y}\,\frac{\partial}{\partial y}\rho(y) \tag{5.9b}$$

provided that

$$\frac{\partial}{\partial y}\rho(y)\underset{\substack{y\to 0\\ y\to\infty}}{\to}0. \tag{5.10}$$

$\mathbf{I}$ in (5.8) is the unit dyadic, and B in (5.9) is a quantity which will be related to the correlation length of the random inhomogeneities at the end of this subsection.

The second dyadic diffusion coefficient can be rewritten as follows:

$$\mathbf{D}_{px}^{rr}=-\int_0^\infty d\tau\,E\left\{\frac{\partial}{\partial\mathbf{x}}\delta\omega_r(\mathbf{x},\mathrm{p};\alpha)\,\frac{\partial}{\partial\mathrm{p}}\delta\omega_r(\mathbf{x}-\mathrm{v}\tau,\mathrm{p};\alpha)\right\}. \tag{5.11}$$

Bearing in mind (5.4), (5.6a), and (5.7), this becomes

$$\mathbf{D}_{px}^{rr}=-g_r(\mathrm{p})\,\frac{\partial}{\partial\mathrm{p}}g_r(\mathrm{p})\,\frac{\mathrm{v}(\mathrm{p})}{v^2(\mathrm{p})}\,\Gamma_{rr}(0)\int_0^\infty dy\,\frac{\partial}{\partial y}\rho(y), \tag{5.12}$$

which, upon integration, changes to

$$\mathbf{D}_{px}^{rr}=-g_r(\mathrm{p})\,\frac{\partial}{\partial\mathrm{p}}g_r(\mathrm{p})\,\frac{\mathrm{v}(\mathrm{p})}{v^2(\mathrm{p})}\,\Gamma_{rr}(0) \tag{5.13}$$

if, in addition to (5.10), one specifies that $\rho(y)\to 0$ as $y\to\infty$.

By virtue of the definition (5.2a), it is easily seen that

$$\mathbf{D}_{xp}^{rr}=-\mathbf{D}_{px}^{rr}. \tag{5.14}$$

It also develops that the last diffusion coefficient is given by

$$\mathbf{D}_{xx}^{rr}=\frac{1}{v(\mathrm{p})}\left(\frac{\partial}{\partial\mathrm{p}}g_r(\mathrm{p})\right)^2\Gamma_{rr}(0)C, \tag{5.15a}$$

$$C=\int_0^\infty dy\,\rho(y). \tag{5.15b}$$

A physical interpretation of C will be presented later on in this subsection.

We shall turn next to the evaluation of the friction coefficients. The first one [cf. Eq. (5.2e)] is rewritten as

$$\mathbf{F}_p^{ri}=\int_0^\infty d\tau\,E\left\{\frac{\partial}{\partial\mathbf{x}}\delta\omega_r(\mathbf{x},\mathrm{p};\alpha)\,2\delta\omega_i(\mathbf{x}-\mathrm{v}\tau,p;\alpha)\right\}. \tag{5.16}$$

Using (5.4), (5.5), and (5.6b), and (5.7), this yields

$$\mathbf{F}_p^{ri}=-2g_r(\mathrm{p})g_i(\mathrm{p})\,\frac{\mathrm{v}(\mathrm{p})}{v^2(\mathrm{p})}\,\Gamma_{ri}(0). \tag{5.17}$$

An examination of (5.2g) shows that

$$\mathbf{F}_p^{ir}=-\mathbf{F}_p^{ri}. \tag{5.18}$$

The remaining two friction coefficients are found to be equal:

$$\mathbf{F}_x^{ri}=\mathbf{F}_x^{ir}=2g_i(\mathrm{p})\,\frac{\partial}{\partial\mathrm{p}}g_r(\mathrm{p})\,\frac{1}{v(\mathrm{p})}\,\Gamma_{ri}(0)C. \tag{5.19}$$

Finally, the scalar absorption coefficients, A^{ii} and A^i, are given simply by

$$A^{ii}=4g_i^2(\mathrm{p})\,\frac{1}{v(\mathrm{p})}\,\Gamma_{ii}(0)C \tag{5.20}$$

and

$$A^i=2E\{\omega_i(\mathbf{x},\mathrm{p};\alpha)\}. \tag{5.21}$$

The parameters B and C, introduced earlier in this subsection as special integrals of the correlation coefficient, can be considered as measures of the correlation distance of the random processes $\delta\chi_r(\mathbf{x};\alpha)$ and $\delta\chi_i(\mathbf{x};\alpha)$, without any reference to specialized correlation functions. Such a general definition reduces considerably the mathematical complexity of having to work with specific correlation functions chosen from within an already plethoric set of physically meaningful ones, without at the same time detracting much from the physical content of the ensuing results. The motivation for this interpretation is given in the first part of Appendix A.

In the following, we shall use the convention

$$B=1/l,\quad C=\lambda. \tag{5.22}$$

The parameters l and λ will be referred to in the sequel as the *correlation lengths* of the random process. An interpretation of these quantities in terms of the spectral correlation function is provided in the second part of Appendix A.

We now summarize the main results of this section. Introducing (5.9a), (5.13), (5.14), and (5.17)–(5.22) in (5.1), we obtain the general *Fokker–Planck equation*

$$\begin{aligned}
&\left(\frac{\partial}{\partial t}+\mathrm{v}(\mathrm{p})\circ\frac{\partial}{\partial\mathbf{x}}\right)E\{\hat f(\mathbf{x},\mathrm{p},t;\alpha)\}\\
&=\frac{\partial}{\partial\mathrm{p}}\cdot\left[g_r^2(\mathrm{p})\,\frac{\Gamma_{rr}(0)}{l}\,\frac{1}{v(\mathrm{p})}\left(\mathbf{I}-\frac{\mathrm{vv}}{v^2}\right)\cdot\frac{\partial}{\partial\mathrm{p}}E\{\hat f(\mathbf{x},\mathrm{p},t;\alpha)\}\right]\\
&\quad+\frac{\partial}{\partial\mathrm{p}}\circ\left(g_r(\mathrm{p})\,\frac{\partial}{\partial\mathrm{p}}g_r(\mathrm{p})\Gamma_{rr}(0)\,\frac{\mathrm{v}(\mathrm{p})}{v^2(\mathrm{p})}\circ\frac{\partial}{\partial\mathbf{x}}E\{\hat f(\mathbf{x},\mathrm{p},t;\alpha)\}\right)\\
&\quad-\frac{\partial}{\partial\mathbf{x}}\circ\left(g_r(\mathrm{p})\,\frac{\partial}{\partial\mathrm{p}}g_r(\mathrm{p})\Gamma_{rr}(0)\,\frac{\mathrm{v}(\mathrm{p})}{v^2(\mathrm{p})}\circ\frac{\partial}{\partial\mathrm{p}}E\{\hat f(\mathbf{x},\mathrm{p},t;\alpha)\}\right)\\
&\quad+\frac{\partial}{\partial\mathbf{x}}\circ\left[\left(\frac{\partial g_r(\mathrm{p})}{\partial\mathrm{p}}\right)^2\Gamma_{rr}(0)\,\frac{\lambda}{v(\mathrm{p})}\circ\frac{\partial}{\partial\mathbf{x}}E\{\hat f(\mathbf{x},\mathrm{p},t;\alpha)\}\right]\\
&\quad-\frac{\partial}{\partial\mathrm{p}}\cdot\left(2g_r(\mathrm{p})g_i(\mathrm{p})\Gamma_{ri}(0)\,\frac{\mathrm{v}(\mathrm{p})}{v^2(\mathrm{p})}\,E\{\hat f(\mathbf{x},\mathrm{p},t;\alpha)\}\right)\\
&\quad-\frac{\partial}{\partial\mathbf{x}}\cdot\left(2g_i(\mathrm{p})\,\frac{\partial}{\partial\mathrm{p}}g_r(\mathrm{p})\Gamma_{ri}(0)\,\frac{\lambda}{v(\mathrm{p})}\,E\{\hat f(\mathbf{x},\mathrm{p},t;\alpha)\}\right.\\
&\quad+2g_r(\mathrm{p})g_i(\mathrm{p})\Gamma_{ir}(0)\,\frac{\mathrm{v}(\mathrm{p})}{v^2(\mathrm{p})}\cdot\frac{\partial}{\partial\mathrm{p}}E\{\hat f(\mathbf{x},\mathrm{p},t;\alpha)\}\\
&\quad-2g_i(\mathrm{p})\,\frac{\partial}{\partial\mathrm{p}}g_r(\mathrm{p})\Gamma_{ir}(0)\,\frac{\lambda}{v(\mathrm{p})}\cdot\frac{\partial}{\partial\mathbf{x}}E\{\hat f(\mathbf{x},\mathrm{p},t;\alpha)\}\\
&\quad+4g_i^2(\mathrm{p})\Gamma_{ii}(0)\,\frac{\lambda}{v(\mathrm{p})}E\{\hat f(\mathbf{x},\mathrm{p},t;\alpha)\}\\
&\quad+2E\{\omega_i(\mathbf{x},\mathrm{p};\alpha)\}E\{\hat f(\mathbf{x},\mathrm{p},t;\alpha)\}.
\end{aligned} \tag{5.23}$$

This equation should be augmented by the initial mean distribution function $E\{\hat f(\mathbf{x},\mathrm{p},0;\alpha)\}$.

VI. SPECIAL CASES OF THE GENERAL FOKKER-PLANCK EQUATION

We present in this section two simplifications of the

general Fokker–Planck equation (5.23) corresponding to spatially homogeneous system and isotropic media.

A. Spatially homogeneous systems

Besides the assumptions (i) and (ii) made in the previous section we must also impose in this case the condition

$$\frac{\partial}{\partial \mathbf{x}} E\{\hat{f}(\mathbf{x}, \mathbf{p}, t; \alpha)\} = 0. \tag{6.1}$$

It follows, then, that the only nonvanishing of the coefficients (5.2) are $\mathbf{D}_{pp}^{rr}$, $\mathbf{F}_p^{ri}$, $\mathbf{F}_p^{ir}$, A^{ii}, A^i, and (5.23) reduces to the following relaxation equation in momentum space:

$$\begin{aligned}
&\frac{\partial}{\partial t} E\{\hat{f}(\mathbf{p}, t; \alpha)\} \\
&= \frac{\partial}{\partial \mathbf{p}} \cdot \left[g_r^2(\mathbf{p}) \frac{\Gamma_{rr}(0)}{l} \frac{1}{v(\mathbf{p})} \left(\mathbf{I} - \frac{\mathbf{vv}}{v^2} \right) \cdot \frac{\partial}{\partial \mathbf{p}} E\{\hat{f}(\mathbf{p}, t; \alpha)\} \right] \\
&\quad - \frac{\partial}{\partial \mathbf{p}} \cdot \left(2 g_r(\mathbf{p}) g_i(\mathbf{p}) \Gamma_{ri}(0) \frac{\mathbf{v}(\mathbf{p})}{v^2(\mathbf{p})} E\{\hat{f}(\mathbf{p}, t; \alpha)\} \right) \\
&\quad + 2 g_r(\mathbf{p}) g_i(\mathbf{p}) \Gamma_{ir}(0) \frac{\mathbf{v}(\mathbf{p})}{v^2(\mathbf{p})} \cdot \frac{\partial}{\partial \mathbf{p}} E\{\hat{f}(\mathbf{p}, t; \alpha)\} \\
&\quad + 4 g_i^2(\mathbf{p}) \Gamma_{ii}(0) \frac{\lambda}{v(\mathbf{p})} E\{\hat{f}(\mathbf{p}, t; \alpha)\} \\
&\quad + 2 E\{\omega_i(\mathbf{x}, \mathbf{p}; \alpha)\} E\{\hat{f}(\mathbf{p}, t; \alpha)\}.
\end{aligned} \tag{6.2}$$

When written in spherical coordinates in p-space, (6.2) is a generalization of the Fokker–Planck equation obtained by Chernov[6] for the probability, $P(\theta, \phi, s)$, of ray directions (θ, ϕ are spherical coordinates) of arc length s in the Markovian approximation. It is also related to the expression for the coherent distribution function, $E\{f(\mathbf{v}, t; \alpha)\}$, in the problem of stochastic acceleration of uniformly distributed particles under the action of time-independent electric and magnetic fields.

B. Isotropic, dissipative and/or dispersive systems[25]

As a consequence of the isotropy of the medium, $\mathbf{v}(\mathbf{p}) = v(p)\hat{\mathbf{p}}$, $g_r(\mathbf{p}) = g_r(p)$, $g_i(\mathbf{p}) = g_i(p)$, $(\partial/\partial \mathbf{p}) g_r(\mathbf{p}) = (\partial/\partial p) g_r(p) \hat{\mathbf{p}}$, and $E\{\omega_i(\mathbf{x}, \mathbf{p}; \alpha)\} = E\{\omega_i(\mathbf{x}, p; \alpha)\}$, where $\hat{\mathbf{p}} = \mathbf{p}/|\mathbf{p}|$. Under these conditions, the general Fokker–Planck equation (5.23) simplifies to

$$\begin{aligned}
&\left(\frac{\partial}{\partial t} + V(p) \hat{\mathbf{p}} \cdot \frac{\partial}{\partial \mathbf{x}} \right) E\{\hat{f}(\mathbf{x}, \mathbf{p}, t; \alpha)\} \\
&= D_x(p) \left(\hat{\mathbf{p}} \cdot \frac{\partial}{\partial \mathbf{x}} \right)^2 E\{\hat{f}(\mathbf{x}, \mathbf{p}, t; \alpha)\} \\
&\quad + D_p(p) \left(\mathbf{p} \times \frac{\partial}{\partial \mathbf{p}} \right)^2 E\{\hat{f}(\mathbf{x}, \mathbf{p}, t; \alpha)\} \\
&\quad + 2 \nu_{\mathrm{eff}}(p) E\{\hat{f}(\mathbf{x}, \mathbf{p}, t; \alpha)\},
\end{aligned} \tag{6.3a}$$

with

$$\begin{aligned}
V(p) &= v(p) + \frac{1}{2} \frac{\partial}{\partial p} \left[\left(\frac{\partial}{\partial p} g_r^2(p) \right) \frac{1}{v(p)} \right] \Gamma_{rr}(0) \\
&\quad + \frac{1}{2} \left(\frac{\partial}{\partial p} g_r^2(p) \right) \frac{1}{v(p)} \Gamma_{rr}(0) \frac{\partial}{\partial \mathbf{p}} \cdot \hat{\mathbf{p}}
\end{aligned} \tag{6.3b}$$

$$\begin{aligned}
&- 4 g_i(p) \left(\frac{\partial}{\partial p} g_r(p) \right) \frac{\lambda}{v(p)} \Gamma_{ri}(0), \\
D_x(p) &= \left(\frac{\partial}{\partial p} g_r(p) \right)^2 \frac{\lambda}{v(p)} \Gamma_{rr}(0),
\end{aligned} \tag{6.3c}$$

$$D_p(p) = \frac{1}{p^2} \frac{g_r^2(p)}{v(p)} \frac{\Gamma_{rr}(0)}{l}, \tag{6.3d}$$

$$\begin{aligned}
\nu_{\mathrm{eff}}(p) &= E\{\omega_i(\mathbf{x}, p; \alpha)\} + \frac{1}{2} \left\{ - \frac{\partial}{\partial p} \left(2 g_r(p) g_i(p) \frac{1}{v(p)} \right) \Gamma_{ri}(0) \right. \\
&\quad - \left(2 g_r(p) g_i(p) \frac{1}{v(p)} \right) \Gamma_{ri}(0) \frac{\partial}{\partial \mathbf{p}} \cdot \hat{\mathbf{p}} \\
&\quad \left. + 4 g_i^2(p) \frac{\lambda}{v(p)} \Gamma_{ii}(0) \right\}.
\end{aligned} \tag{6.3e}$$

In (6.3a), the left-hand side describes convection with modified group velocity $V(p)$, the first term on the right-hand side describes longitudinal spatial diffusion, the second term arises because of angular diffusion in momentum space, and, finally, the last term designates the effective absorption in the random medium.

In Appendix B we shall discuss the specific forms of Eq. (6.3a), as well as the coefficients $V(p)$, $D_x(p)$, $D_p(p)$, and $\nu_{\mathrm{eff}}(p)$, for three-, two-, and one-dimensional problems.

APPENDIX A: PHYSICAL INTERPRETATION OF THE PARAMETERS l AND λ

In order to motivate our interpretation in Sec. 5 of the quantities l and λ as general measures of the correlation length of the random process, without recourse to specialized correlation functions, we refer to two special, but widely used correlation coefficients:

$$\text{(i)}\ \rho(y) = \exp(-y^2/L^2), \quad l = L/\sqrt{\pi}, \quad \lambda = (\sqrt{\pi}/2)L, \tag{A1}$$

$$\text{(ii)}\ \rho(y) = [1 + (y/L)^2]^{-2}, \quad l = (4/3\pi)L, \quad \lambda = (\pi/8)L. \tag{A2}$$

In these two examples, L denotes the correlation length of the random process under consideration.

In the second part of this appendix, we provide a physical interpretation of the parameters l and λ with the aid of the spectral correlation function.

In Sec. 5, the correlation lengths l and λ were defined in terms of the correlation coefficient $\rho(y)$ as follows:

$$\frac{1}{l} = -\int_0^\infty dy \frac{1}{y} \frac{\partial}{\partial y} \rho(y), \tag{A3}$$

$$\lambda = \int^\infty dy\, \rho(y). \tag{A4}$$

The spectral correlation function is introduced next by means of the integral

$$S(p) = (2\pi)^{-3} \int_{R^3} d\mathbf{y}\, \rho(y) \exp(-i\mathbf{p} \cdot \mathbf{y}) \tag{A5}$$

Integrating (A5) over p yields

$$\int_{R^3} d\mathbf{p} S(p) = 1 \tag{A6}$$

since $\rho(0) = 1$. An inversion of the Fourier transformation (A5) results in the expression

$$\rho(y) = (4\pi/y) \int_0^\infty dp\, p\, (\sin py) S(p). \tag{A7}$$

Introducing this result in (A3) and (A4) gives rise to the relations

$$\frac{1}{l} = -4\pi \int_0^\infty dp\, pS(p) \left[\int_0^\infty dy \frac{1}{y} \frac{\partial}{\partial y} \left(\frac{\sin py}{y} \right) \right], \tag{A8}$$

$$\lambda = 4\pi \int_0^\infty dp\, pS(p) \left[\int_0^\infty dy \left(\frac{\sin py}{y} \right) \right], \tag{A9}$$

However, the definite integrals over y appearing in (A8) and (A9) can be carried out explicitly, viz.,

$$\int_0^\infty dy \frac{1}{y} \frac{\partial}{\partial y} \left(\frac{\sin py}{y} \right) = -\frac{\pi}{4} p^2, \tag{A10}$$

$$\int_0^\infty dy \frac{\sin py}{y} = \frac{\pi}{2}. \tag{A11}$$

Therefore, one finally has

$$1/l = (\pi/4) \int_0^\infty dp\, p^3 S(p), \tag{A12}$$

$$\lambda = (\pi/2) \int_0^\infty dp\, pS(p). \tag{A13}$$

Similar expressions can be written for the two-dimensional and the one-dimensional case.

APPENDIX B: THREE-, TWO-, AND ONE-DIMENSIONAL FOKKER-PLANCK EQUATIONS

When the Fokker—Planck equation (6.3a) for an isotropic, dissipative and/or dispersive medium is considered in a three-dimensional Euclidean space, it is convenient to introduce a spherical polar coordinate system in momentum space: $\mathbf{p} = (p, \theta, \phi)$. Then we use (6.3a) with

$$\left(\mathbf{p} \times \frac{\partial}{\partial \mathbf{p}} \right)^2 = \frac{1}{\sin^2\theta} \frac{\partial^2}{\partial \phi^2} + \frac{1}{\sin\theta} \frac{\partial}{\partial \theta} \left(\sin\theta \frac{\partial}{\partial \theta} \right). \tag{B1}$$

The coefficients $V(p)$, $D_x(p), \cdots$, are given as in (6.3b), (6.3c), $\cdots$, with $(\partial/\partial \mathbf{p}) \cdot \hat{\mathbf{p}} = (2/p)$.

In the two-dimensional case, we introduce a polar coordinate system in momentum space: $\mathbf{p} = (p, \phi)$. Then, in examining specific problems, we must use (6.3a) with

$$\left(\mathbf{p} \times \frac{\partial}{\partial \mathbf{p}} \right)^2 = \frac{\partial^2}{\partial \phi^2}, \quad \frac{\partial}{\partial \mathbf{p}} \cdot \hat{\mathbf{p}} = \frac{1}{p}. \tag{B2}$$

In the one-dimensional case one has

$$\left(\mathbf{p} \times \frac{\partial}{\partial \mathbf{p}} \right)^2 = 0, \quad \frac{\partial}{\partial \mathbf{p}} \cdot \hat{\mathbf{p}} = 0. \tag{B3}$$

Equation (6.3a) assumes the simpler form

$$\left(\frac{\partial}{\partial t} + V(p) \frac{\partial}{\partial x} \right) E\{\hat{f}(x,p,t;\alpha)\}$$
$$= D_x(p) \frac{\partial^2}{\partial x^2} E\{\hat{f}(x,p,t;\alpha)\} \tag{B4a}$$
$$+ 2\nu_{\text{eff}}(p) E\{\hat{f}(x,p,t;\alpha)\},$$

and the coefficients $V(p)$, $D_x(p)$, and $\nu_{\text{eff}}(p)$ are modified as follows:

$$V(p) = v(p) - \frac{1}{2} \left\{ \frac{\partial}{\partial p} \left(\frac{\partial}{\partial p} g_r^2(p) \frac{1}{v(p)} \right) \Gamma_{rr}(0) \right. \tag{B4b}$$
$$\left. - 4 g_i(p) \frac{\partial}{\partial p} g_r(p) \frac{\lambda}{v(p)} \Gamma_{ri}(0) \right\},$$

$$D_x(p) = \left(\frac{\partial}{\partial p} g_r(p) \right)^2 \frac{\lambda}{v(p)} \Gamma_{rr}(0), \tag{B4c}$$

$$\nu_{\text{eff}}(p) = E\{\omega_i(x,p;\alpha)\}$$
$$+ \left\{ -\frac{\partial}{\partial p} \left(2 g_r(p) g_i(p) \frac{1}{v(p)} \right) \Gamma_{ri}(0) \right. \tag{B4d}$$
$$\left. + 4 g_i^2(p) \frac{\lambda}{v(p)} \Gamma_{ii}(0) \right\}.$$

The transport equation (B4) has been used to study the problem of wave packet spreading on a random transmission line (cf. Ref. 3) and the propagation of frequency-modulated pulses in a randomly stratified plasma (cf. Ref. 4).

*The research reported in this paper was completed while the authors participated in the Applied Mathematics Summer Institute, 1975 at Dartmouth College. The Institute was supported by the Office of Naval Research under Contract No. N0014-75-C-0121 with the Applied Institute of Mathematics, Inc.

†Research supported in part by Contract AFOSR-76-2881.

[1] F. D. Tappert, SIAM Rev. (Chronicle) **13**, 281 (1971).

[2] F. D. Tappert and I. M. Besieris, "Stochastic Wave Kinetic Equation and its Application to Wave Packet Spreading," URSI Intern. Symp. on Electromagnetic Wave Theory, Tbilisi, USSR, Izdatel'stro Nauka, Sibirskoe Otdelenie, Novosibirsk, SSR, pp. 230—34 (1971).

[3] I. M. Besieris and F. D. Tappert, J. Appl. Phys. **44**, 2119 (1973).

[4] I. M. Besieris and F. D. Tappert, J. Math. Phys. **14**, 704 (1973).

[5] I. M. Besieris and F. D. Tappert, J. Math. Phys. **14**, 1829 (1973).

[6] H. Bremmer, Radio Sci. **8**, 511 (1973).

[7] V. Y. Kharanen, Dokl. Akad. Nauk SSSR **88**, 253 (1953).

[8] L. A. Chernov, *Wave Propagation in a Random Medium* (McGraw-Hill, New York, 1960).

[9] J. B. Keller, Proc. Symp. Appl. Math. **13**, 227 (1962).

[10] U. Frisch, "Wave Propagation in Random Media," in *Probabilistic Methods in Applied Mathematics*, Vol. I, edited by A. T. Bharucha-Reid (Academic, New York, 1968).

[11] In the geometric, ray-optical approach to random media, a linear, Liouville-type, partial differential equation is set up for $P(\mathbf{x}, \mathbf{u}, s) = \delta(\mathbf{x} - \mathbf{x}(s))\, \delta(\mathbf{u} - \mathbf{u}(s))$, where $\mathbf{x}(s)$ and $\mathbf{u}(s)$ denote respectively the position and direction of a ray of arclength s.

The ensemble average of $P(\mathbf{x},\mathbf{u},s)$ gives the simultaneous probability of position and ray direction.
[12]R. Kubo, J. Math. Phys. **4**, 174 (1963); S. Orszag and R. Kraichnan, Phys. Fluids **10**, 1720 (1967); A.N. Kaufman, Phys. Fluids **11**, 326 (1968); E.C. Roelof, Can. J. Phys. **46**, S990 (1969); C.E. Newman, P.A. Sturrock, and E. Tademaru, Astrophys. Space Sci. **10**, 102 (1971); G. Papanicolau, J. Math. Phys. **12**, 1493 (1971); J.R. Jokipii, Astrophys. J. 172, 319 (1972).
[13]R.M. Lewis, Arch. Ratl. Mech. Anal. **20**, 191 (1965).
[14]E. Wigner, Phys. Rev. **40**, 749 (1932).
[15]J.E. Moyal, Proc. Cambridge Phil. Soc. **45**, 99 (1949); B. Leaf, J. Math. Phys. **9**, 65, 769 (1968).
[16]The space and momentum coordinates **x**, **p**, as well as the parameter α appearing in the argument of the distribution function will occasionally be suppressed for convenience in this section.
[17]R. Balescu, Physica **38**, 98 (1968); **42**, 464 (1969).
[18]I.M. Besieris, J. Math. Phys. **13**, 358 (1972).
[19]J. Weinstock, Phys. Fluids **12**, 1045 (1969); **13**, 2308 (1970).
[20]R. Balescu and J.H. Misguich, J. Plasma Phys. **11**, 357, 377 (1974); **13**, 33, 53 (1974).
[21]W.C. Meecham, "On Radiation in a Randomly Inhomogeneous Medium," Space Tech. Lab. Rpt. BSD-TR-61-36, Los Angeles, California (1961).
[22]R.C. Bourret, Nuovo Cimento **26**, 1 (1962); Can. J. Phys. **40**, 782 (1962).
[23]J.B. Keller, Proc. Symp. Appl. Math. **16**, 145 (1964).
[24]In writing down (3.12), it has been assumed that the operator L_1 has zero mean.
[25]The main results in this section have been used in Ref. 2 in order to examine the problem of spreading of planar wave-packets having Gaussian envelopes.

Monotonicity of correlation functions

Paul A. Pearce

Mathematics Department, University of Melbourne, Parkville, Victoria 3052, Australia
(Received 14 August 1975)

Counterexamples to pair correlation monotonicity inequalities, analogous to Griffiths' second inequality for the Ising model, are presented for the finite spin Heisenberg model, the spin-(1/2) X–Y model, the anisotropic planar classical Heisenberg model, and the spherical model.

1. INTRODUCTION

Griffiths[1] has shown that for Ising ferromagnets the spin correlations are nonnegative and monotonic increasing functions of the interactions. The natural extension to the isotropic planar classical Heisenberg model was given by Ginibre.[2] Similar theorems have also been proven for the spin-$\frac{1}{2}$ X–Y model[3] and more recently for the anisotropic planar classical Heisenberg model[4] regarding spin *component* correlations.

No analog of Griffiths' second theorem has been found for a ferromagnet with a three-dimensional order parameter. Indeed, Hurst and Sherman[5] have shown that for a spin-$\frac{1}{2}$ Heisenberg chain the pair correlations are not, in general, monotonic increasing functions of the interactions. Yeh[6] subsequently observed that for a classical (spin-∞) Heisenberg chain the pair correlations are independent of the interactions between other pairs. The possibility of regaining Griffiths' second theorem at sufficiently high spin values is dispelled in Sec. 2 where it is shown that the Hurst and Sherman counterexample holds for all finite spin values. The important case of the classical (spin-∞) Heisenberg model remains unresolved with little hope for a simple counterexample since, in this case, the theorem is true at least for the chain-type structures[6] and Husimi trees.[7]

In Sec. 3 it is shown that the pair correlations are also not monotonic increasing functions of the interactions for the spin-$\frac{1}{2}$ X–Y model. This supports the view that quantum effects are responsible for the loss of monotonicity of the correlations. For higher spin values the simple counterexample presented breaks down. This matter, however, is not pursued further here.

A particular monotonicity inequality obtained by Ginibre[2] for the isotropic planar classical Heisenberg model is of the form

$$\langle(\mathbf{S}_i\cdot\mathbf{S}_j)(\mathbf{S}_k\cdot\mathbf{S}_l)\rangle-\langle\mathbf{S}_i\cdot\mathbf{S}_j\rangle\langle\mathbf{S}_k\cdot\mathbf{S}_l\rangle\geq 0. \tag{1.1}$$

In Sec. 4 it is shown that this inequality is no longer true when spin space anisotropy is introduced.

Finally in Sec. 5 a counterexample is presented to Griffiths-type inequalities for the Berlin and Kac[8] spherical model.

2. HEISENBERG MODEL

Consider a system of three Heisenberg spins with Hamiltonian

$$\mathcal{H}=-s^2(J_1\mathbf{S}_1\cdot\mathbf{S}_2+J_2\mathbf{S}_2\cdot\mathbf{S}_3), \tag{2.1}$$

where S_i, $i=1,2,3$, is the spin-s operator for the ith site. For weak interactions (high temperatures) the logarithm of the partition function

$$Z=\mathrm{Tr}\exp(-\mathcal{H}), \tag{2.2}$$

is given by the cumulant expansion:

$$\begin{aligned}\log Z=\log\mathrm{Tr}1&-\langle\mathcal{H}\rangle+(1/2!)[\langle\mathcal{H}^2\rangle-\langle\mathcal{H}\rangle^2]\\&-(1/3!)[\langle\mathcal{H}^3\rangle-3\langle\mathcal{H}^2\rangle\langle\mathcal{H}\rangle+2\langle\mathcal{H}\rangle^3]\\&+(1/4!)[\langle\mathcal{H}^4\rangle-4\langle\mathcal{H}^3\rangle\langle\mathcal{H}\rangle-3\langle\mathcal{H}^2\rangle^2+12\langle\mathcal{H}^2\rangle\langle\mathcal{H}\rangle^2-6\langle\mathcal{H}\rangle^4]\\&-\cdots,\end{aligned} \tag{2.3}$$

where

$$\langle\cdots\rangle=\mathrm{Tr}(\cdots)/\mathrm{Tr}1. \tag{2.4}$$

From the cyclic invariance of the trace

$$\frac{\partial}{\partial J_1}\langle\mathbf{S}_2\cdot\mathbf{S}_3\rangle=\frac{\partial}{\partial J_1}\left(Z^{-1}\frac{\partial Z}{\partial J_2}\right)=\frac{\partial^2}{\partial J_1\partial J_2}\log Z. \tag{2.5}$$

The differentiation of the expansion for $\log Z$ is simplified by observing that $\mathrm{Tr}S_i=0$, $i=1,2,3$, leads to the consequences:

$$\langle\mathcal{H}\rangle=\langle\mathcal{H}^3\rangle=0$$

and (2.6)

$$\frac{\partial^2}{\partial J_1\partial J_2}\langle\mathcal{H}^2\rangle=0.$$

Clearly the lowest order contribution to (2.5) is

$$\frac{1}{4!}\frac{\partial^2}{\partial J_1\partial J_2}\{\langle\mathcal{H}^4\rangle-3\langle\mathcal{H}^2\rangle^2\}, \tag{2.7}$$

which after elementary manipulations becomes

$$\begin{aligned}&\tfrac{1}{3}s^{-8}J_1J_2[2\langle(\mathbf{S}_1\cdot\mathbf{S}_2)^2(\mathbf{S}_2\cdot\mathbf{S}_3)^2\rangle\\&+\langle(\mathbf{S}_1\cdot\mathbf{S}_2)(\mathbf{S}_2\cdot\mathbf{S}_3)(\mathbf{S}_1\cdot\mathbf{S}_2)(\mathbf{S}_2\cdot\mathbf{S}_3)\rangle\\&-3\langle(\mathbf{S}_1\cdot\mathbf{S}_2)^2\rangle^2\}.\end{aligned} \tag{2.8}$$

The traces appearing in (2.8) have been evaluated as polynomials in $\eta=s(s+1)$ by Subramanian and Devanathan.[9] Using their results, expression (2.8) becomes

$$\begin{aligned}&\tfrac{1}{3}s^{-8}J_1J_2\{\tfrac{1}{30}\eta^3[2(2\eta+1)+\tfrac{2}{3}(4\eta-3)+\tfrac{1}{3}(4\eta+2)+2(\eta-2)]-\tfrac{1}{3}\eta^4\}\\&=-\tfrac{1}{27}s^{-8}\eta^3J_1J_2.\end{aligned} \tag{2.9}$$

Hence, for any finite spin,

$$\frac{\partial}{\partial J_1}\langle \mathbf{S}_2 \cdot \mathbf{S}_3\rangle = -\frac{1}{27}s^{-8}\eta^3 J_1 J_2 + \cdots < 0, \tag{2.10}$$

for J_1 and J_2 sufficiently small and positive. Note that the term on the right-hand side of (2.10) vanishes as s tends to infinity in accordance with Ginibre. Also, $s=\frac{1}{2}$ implies $\eta=\frac{3}{4}$ and $\frac{1}{27}s^{-8}\eta^3=4$, in agreement with Hurst and Sherman.

3. SPIN-½ X-Y MODEL

Consider a system of three X—Y spins with Hamiltonian

$$\mathcal{H}_{X\text{-}Y} = -s^{-2}(J_1 \mathbf{s}_1 \cdot \mathbf{s}_2 + J_2 \mathbf{s}_2 \cdot \mathbf{s}_3), \tag{3.1}$$

where $\mathbf{s}_i$, $i=1,2,3$, is the projection of the spin-s operator onto the X—Y plane. By the analysis of Sec. 2, the lowest order contribution to $(\partial/\partial J_1)\langle \mathbf{s}_2 \cdot \mathbf{s}_3\rangle$ for weak interactions is

$$\tfrac{1}{3}s^{-8}J_1J_2\{2\langle(\mathbf{s}_1\cdot\mathbf{s}_2)^2(\mathbf{s}_2\cdot\mathbf{s}_3)^2\rangle + \langle(\mathbf{s}_1\cdot\mathbf{s}_2)(\mathbf{s}_2\cdot\mathbf{s}_3)(\mathbf{s}_1\cdot\mathbf{s}_2)(\mathbf{s}_2\cdot\mathbf{s}_3)\rangle - 3\langle(\mathbf{s}_1\cdot\mathbf{s}_2)^2\rangle^2\} \tag{3.2}$$

$$= \tfrac{1}{3}s^{-8}J_1J_2\{\tfrac{1}{30}\eta[2(2\eta+1)(\tfrac{2}{3}\eta)^2 + 2(4\eta-3)(\tfrac{1}{3}\eta)(\tfrac{2}{3}\eta) + (4\eta+2)(\tfrac{1}{3}\eta)(\tfrac{2}{3}\eta) + 2(\eta-2)(\tfrac{2}{3}\eta)^2] - 3(\tfrac{2}{9}\eta^2)^2\}$$

$$= \tfrac{4}{405}s^{-8}\eta^3(\eta-2)J_1J_2. \tag{3.3}$$

Hence

$$\frac{\partial}{\partial J_1}\langle \mathbf{s}_1 \cdot \mathbf{s}_2\rangle = \tfrac{4}{405}s^{-8}\eta^3(\eta-2)J_1J_2 + \cdots < 0, \tag{3.4}$$

for $s=\frac{1}{2}$ (i.e., $\eta=\frac{3}{4}$) and for J_1 and J_2 sufficiently small and positive. The counterexample fails for higher spin values (i.e., $\eta \geq 2$).

4. ANISOTROPIC PLANAR CLASSICAL HEISENBERG MODEL

Consider an anisotropic planar classical Heisenberg chain described by the Hamiltonian

$$\mathcal{H}_{\text{Anis}} = (J_1+K_1)S_{1x}S_{2x} + (J_2+K_2)S_{2x}S_{3x} + J_1S_{1y}S_{2y} + J_2S_{2y}S_{3}y \tag{4.1}$$

where $\mathbf{S}_i = (S_{ix}, S_{iy})$, $i=1,2,3$, are unit plane vectors. By similar reasoning to that in Sec. 2 it is seen that the leading term in a high temperature expansion of $(\partial/\partial J_1)\langle \mathbf{S}_2 \cdot \mathbf{S}_3\rangle$ is

$$\frac{1}{4!}\frac{\partial^2}{\partial J_1 \partial J_2}\{\langle \mathcal{H}_{\text{Anis}}{}^4\rangle - 3\langle \mathcal{H}_{\text{Anis}}{}^2\rangle^2\}. \tag{4.2}$$

For this model, the configurational integrals are given by

$$\langle S_x^m S_y^n\rangle = \begin{cases} (m-1)!!\,(n-1)!!/(m+n)!!, & m,n \text{ even}, \\ 0, & \text{otherwise}. \end{cases} \tag{4.3}$$

Consequently, the expansion of the leading term (4.2) gives

$$\frac{1}{4}\frac{\partial^2}{\partial J_1 \partial J_2}\{[(J_1+K_1)^2(J_2+K_2)^2 + J_1^2J_2^2][\langle(S_{1x}S_{2x})^2(S_{2x}S_{3x})^2\rangle - \langle(S_{1x}S_{2x})^2\rangle^2] + [J_2^2(J_1+K_1)^2 + J_1^2(J_2+K_2)^2] \times [\langle(S_{1x}S_{2x})^2(S_{2y}S_{3y})^2\rangle - \langle(S_{1x}S_{2x})^2\rangle^2]\}$$

$$= \tfrac{1}{4}[(J_1+K_1)(J_2+K_2) + J_1J_2]\cdot(\tfrac{3}{8}-\tfrac{1}{4}) + \tfrac{1}{4}[J_2(J_1+K_1) + J_1(J_2+K_2)]\cdot(\tfrac{1}{8}-\tfrac{1}{4}) = \tfrac{1}{32}K_1K_2. \tag{4.4}$$

Hence,

$$\langle(\mathbf{S}_1\cdot\mathbf{S}_2)(\mathbf{S}_2\cdot\mathbf{S}_3)\rangle - \langle \mathbf{S}_1\cdot\mathbf{S}_2\rangle\langle \mathbf{S}_2\cdot\mathbf{S}_3\rangle = \tfrac{1}{32}K_1K_2 + \cdots < 0, \tag{4.5}$$

for sufficiently high temperatures if the perturbation parameters K_1 and K_2 are chosen so that $K_1K_2<0$, $J_1+K_1>0$, and $J_2+K_2>0$.

5. SPHERICAL MODEL

Consider a spherical model consisting of four spins $\{x_i\}_{i=1}^4$ with "Hamiltonian"

$$\mathcal{H}_{\text{sph}} = -J_{12}x_1x_2 - J_{34}x_3x_4, \tag{5.1}$$

and partition function

$$Z_{\text{sph}} = \int_{\sum_{i=1}^4 x_i^2=4} \cdots \int \prod_{i=1}^4 dx_i \exp(-\mathcal{H}_{\text{sph}}). \tag{5.2}$$

The quadratic form $-\mathcal{H}_{\text{sph}}$ is associated with the symmetric matrix $\frac{1}{2}J$, where

$$J = \begin{bmatrix} 0 & J_{12} & 0 & 0 \\ J_{12} & 0 & 0 & 0 \\ 0 & 0 & 0 & J_{34} \\ 0 & 0 & J_{34} & 0 \end{bmatrix}. \tag{5.3}$$

Scaling the spins and introducing the delta function gives

$$Z_{\text{sph}} = 8\int_{-\infty}^{\infty}\cdots\int \prod_{i=1}^4 dx_i\,\delta\Big(1-\sum_{i=1}^4 x_i^2\Big)\exp(4J_{12}x_1x_2 + 4J_{34}x_3x_4). \tag{5.4}$$

Finally, the substitution of the delta function integral representation

$$\delta(x) = \frac{1}{2\pi i}\int_{-i\infty}^{i\infty}\exp(xs)\,ds \tag{5.5}$$

into (5.4) yields, after interchanging orders of integration, [8]

$$\tfrac{1}{8}Z_{\text{sph}} = \frac{1}{2\pi i}\int_{\alpha-i\infty}^{\alpha+i\infty} e^s \Xi(s)\,ds, \tag{5.6}$$

where

$$\Xi(s) = \int_{-\infty}^{\infty}\cdots\int \prod_{i=1}^4 dx_i \exp\Big(-s\sum_{i=1}^4 x_i^2 + 4J_{12}x_1x_2 + 4J_{34}x_3x_4\Big), \tag{5.7}$$

and $\alpha>0$ is chosen so that the line $\mathrm{re}\, s=\alpha$ is to the right of all singularities of $\Xi(s)$.

The integrations in (5.7) are readily performed by transforming the spin variables so as to diagonalize the

quadratic form $-H_{\rm sph}$. This procedure leads to the result

$$\Xi(s) = \pi^2[\det(sI - 2J)]^{-1/2} \tag{5.8}$$

$$= \pi^2(s^2 - 4J_{12}^2)^{-1/2}(s^2 - 4J_{34}^2)^{-1/2}. \tag{5.9}$$

The partition function can now be evaluated by noticing from (5.6) that

$$(1/8)Z_{\rm sph} = \xi(1), \tag{5.10}$$

where $\xi(t)$ is the inverse Laplace transform

$$\xi(t) = \frac{1}{2\pi i}\int_{\alpha - i\infty}^{\alpha + i\infty} \exp(st)\Xi(s)\,ds. \tag{5.11}$$

Since[10]

$$I_0(at) = \frac{1}{2\pi i}\int_{\alpha - i\infty}^{\alpha + i\infty} \exp(st)(s^2 - a^2)^{-1/2}\,ds, \tag{5.12}$$

it follows from the Laplace convolution formula that

$$Z_{\rm sph} = 8\pi^2 \int_0^1 I_0(2J_{12}t)I_0[2J_{34}(1-t)]\,dt. \tag{5.13}$$

The ascending series for the zero order modified Bessel function[11] is

$$I_0(z) = \sum_{k=0}^{\infty}(z/2k!)^2. \tag{5.14}$$

Hence, for weak interactions

$$Z_{\rm sph} = 8\pi^2 \int_0^1 \{1 + J_{12}^2t^2 + \cdots\}\{1 + J_{34}^2(1-t)^2 + \cdots\}\,dt \tag{5.15}$$

$$= 8\pi^2\{1 + \tfrac{1}{3}(J_{12}^2 + J_{34}^2) + \tfrac{1}{30}J_{12}^2J_{34}^2 + \cdots\}. \tag{5.16}$$

Now the logarithm series

$$\log(1+x) = x - \tfrac{1}{2}x^2 + \tfrac{1}{3}x^3 - \cdots, \tag{5.17}$$

leads to

$$\frac{\partial^2}{\partial J_{12}\partial J_{34}}\log Z_{\rm sph} = 4J_{12}J_{34}\left\{\frac{1}{30} - \frac{1}{2}\cdot\frac{2}{9} + \cdots\right.$$

$$= -\frac{14}{45}J_{12}J_{34} + \cdots < 0, \tag{5.18}$$

for J_{12}, J_{34} sufficiently small and positive.

ACKNOWLEDGMENT

The author is indebted to his supervisor Professor Colin J. Thompson for suggesting the subject matter and for a critical reading of the manuscript.

[1]R. B. Griffiths, J. Math. Phys. 8, 478, 484 (1967).
[2]J. Ginibre, Comm. Math. Phys. 16, 310 (1970).
[3]G. Gallavotti, Stud. Appl. Math. 50, 89 (1971).
[4]J. L. Monroe, J. Math. Phys. 16, 1809 (1975); P. A. Pearce and C. J. Thompson, "Correlation Function Inequalities for the Planar Classical Heisenberg Model," Research Report No. 16 Mathematics Dept., Univ. of Melbourne (1975).
[5]C. A. Hurst and S. Sherman, J. Math. Phys. 11, 2473 (1970).
[6]R. H. T. Yeh, J. Math. Phys. 11, 1317 (1970).
[7]P. A. Pearce and C. J. Thompson, "Griffiths-type Inequalities for Classical Heisenberg Rings," unpublished.
[8]T. H. Berlin and M. Kac, Phys. Rev. 86, 821 (1952).
[9]P. R. Subramanian and V. Devanathan, J. Phys. A 7, 1995 (1974).
[10]A. Erdélyi (ed.), *Tables of Integral Transforms* (McGraw-Hill, New York, 1954), Vol. I., p. 195.
[11]M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions* (National Bureau of Standards, A.M.S. 55, 1964), p. 375.

Rigorous bounds on inclusive and exclusive cross sections. I. Lower bound on the integral of an analytic function*

J. J. G. Scanio and P. Suranyi

Department of Physics, University of Cincinnati, Cincinnati, Ohio 45221
(Received 10 April 1975)

A lower bound on the integral of an analytic function with certain properties is derived. The properties of the function are those expected of an inclusive or exclusive differential cross section. The integral bound can be used to derive rigorous bounds on these cross sections.

I. INTRODUCTION

There has been interest recently in the study of rigorous bounds on inclusive and exclusive cross sections.[1-4] The bounds obtained in Ref. 3 contain no unknown constants and can be tested experimentally, and in addition they are saturated to within logarithmic factors of the energy if the inclusive cross section scales at high energy.

Unfortunately, the method of proof used in Refs. 1–3 relied heavily on a partial wave expansion of the inclusive amplitude and on the use of the Schwarz inequality to reduce the problem to the study of elastic partial wave amplitudes. Such a reduction is not in general possible and the techniques of Refs. 1–3 cannot be used to obtain rigorous bounds on other types of cross sections.

In this paper we prove the following theorem which will be directly applicable to the derivation of rigorous bounds on various types of cross sections.

Theorem: Let $f(z)$ be a function of the complex variable z, with the following properties:

(i) $f(z)$ is analytic within the ellipse $z=\cosh(d+i\theta)$;

(ii) $|f(z)|\leq c$ for z on the ellipse;

(iii) $f(z)$ is real along the real axis inside the ellipse;

(iv) $f(z)$ is nonnegative on the real axis between -1 and $+1$;

(v) $f(z)$ is normalized so that $f(1)=1$.

Then to leading order in $\ln c$ and $\ln(1/d)$

$$\int_{-1}^{1} f(x)\,dx \geq 2d^2[\ln(c^{1/2})]^{-2}. \tag{1}$$

The properties given to $f(z)$ in the statement of the theorem are appropriate if $f(z)$ is to represent either an exclusive or an inclusive cross section. We will mention two examples of the use of the theorem here.

If we let

$$\left(\frac{d\sigma_{el}}{dz}\right)_{z=1} f(z)=\frac{d\sigma_{el}}{dz}$$

where $z=\cos\theta$ is the scattering angle, using the Jin–Martin bound on the magnitude of the scattering amplitude on the ellipse to obtain $c=(s/s_0)^2$ and letting $d^2=16\mu^2/s$ where μ is the pion mass, we have[5]

$$\sigma_{el}(s)\geq\left(\frac{d\sigma_{el}}{dt}\right)_{t=0} 16\mu^2[\ln(s/s_0)]^{-2}. \tag{2}$$

We can also obtain bounds on inclusive cross sections which have already been derived using other methods.[1-4] For example, using the notation of Ref. 3 if we let

$$F(\cos\theta)=2\pi\int dpp^2 f(p\cos\theta, p\sin\theta, s), \tag{3}$$

where $f(p\cos\theta, p\sin\theta, s)$ is the invariant inclusive cross section we obtain

$$F(1)<\sigma_{tot}\,s^{3/2}\,(\ln s)^2(8\mu^2)^{-1} \tag{4}$$

where we have used the normalization on $F(z)$; $\int_{-1}^{1} F(z)\,dz=s^{1/2}\sigma_{tot}$. The result of Eq. (4) illustrates the power of our theorem. The result of Refs. 1–3 has been obtained without any reference to a partial wave expansion. We leave further applications to a future publication.

The proof of the theorem is somewhat involved and will proceed in a number of steps, with each step in general reducing the complexity of the functions we must study in order to obtain a minimal integral. In Sec. II we show that we need only consider those functions for which $|f(z)|=c$ everywhere on the ellipse. In Sec. III we show that the functions giving a minimal integral must have only real double zeros and no complex zeros. We are thus able to consider the square root of the original function without introducing any singularities in the ellipse. In Sec. IV we derive our minimal bound by first showing that we need only consider finite Legendre polynomial expansions of the square root functions and then by estimating the Legendre expansion in terms of the bound on the ellipse. In Sec. V we indicate that the bound cannot be substantially improved and in the Appendix we give a few properties of the Jacobi elliptic sine functions which we need in Sec. III.

II. BOUND ON THE ELLIPSE

In this section we show that the minimal integral in Eq. (1) will occur for functions $f(z)$ such that $|f(z)|=c$ everywhere on the ellipse.

We first give some definitions and then prove a number of lemmas.

The ellipse E_d for any real d is the locus of points $z=\cosh(d+i\theta)$, $0\leq\theta\leq 2\pi$.

The norm of a function f is given by $\|f\|=\int_{-1}^{1} f(x)\,dx$.

A function f is in class K [more specifically class $K(c)$] if: (i) $f(z)$ is analytic inside $E_{d+\delta}$; (ii) $f(z)$ is real on the real axis inside the ellipse; (iii) $f(1)=1$; (iv) $f(z)$ is nonnegative in the interval $(-1,+1)$; (v) $|f(z)|\leq c$ if $z\in E_d$; (vi) $f(z)$ has no zeros between the ellipses $E_{d+\epsilon}$ and $E_{d-\epsilon}$, where $\delta>\epsilon>0$.

The function $f(z)$ is minimal on K if $f \in K$ and $\|f\| \leq \|g\|$ for every $g \in K$.

It is easy to see that if $f(z)$ satisfies conditions (i)—(v) of a class K function then $f(z)$ is of class K. Since $f(z)$ is analytic inside $E_{d+\delta}$ it has a finite number of zeros inside the ellipse $E_{d+(\delta/2)}$. We can then always pick a δ' with $0 < \delta' < \delta/2$ such that the ellipse $E_{d+\delta'}$ has a neighborhood without any zeros.

We now prove a series of lemmas.

Lemma 1: Let $f(z) \in K$ and let $\|f\|$ be minimal on K. Then there are infinitely many points $z_i \in E_d$ where $|f(z_i)| = c$.

Proof: Suppose there are only a finite number of points N (including zero) such that $|f(z_i)| = c$. We show that $f(z)$ cannot be minimal by constructing the function

$$g(z) = \left[1 - \prod_{i=1}^{N} (z - z_i)^n (z - z_i^*)^n (z-1)^2 a\right] f(z), \tag{5}$$

where a and n are to be fixed later. Clearly $g(z)$ satisfies conditions (i)—(iii) of a class K function. If $\theta < a < a_0 = (2\cosh d)^{-2nN-2}$ then $g(x)$ satisfies condition (iv) and $\|g\| < \|f\|$. To show that $|g(z)| \leq c$ on the ellipse first notice that because $f(z)$ is regular and $|f(z_i)| = c$ we can give the bound $|f(z)| \leq c - a_i |z - z_i|^{n_i}$ when z is on the ellipse and $|z - z_i| < \epsilon_i$. If we choose $n = \max_i n_i$ and $a < (a_0/c)\min_i a_i$ then we have from Eq. (5) that $|g(z)| \leq c$ for $|z - z_i| < \epsilon_i$. Outside these intervals we have $|f(z)| < c - \eta$ for some $\eta > 0$. Choosing a so that it also satisfies $|\prod_{i=1}^{N} (z - z_i)^n (z - z_i^*)^n (z-1)^2 ac| < \eta$ for all $z \in E_d$ we see that $g \in K$. However $\|g\| < \|f\|$ contrary to the minimality assumption for $\|f\|$ and therefore we conclude that $|f(z_i)| = c$ for an infinite number of points $z_i \in E_d$.

Because of this we can show Lemma 2.

Lemma 2: If $f \in K$ and $f(z)$ is minimal on K, then $|f(z)| = c$ for all $z \in E_d$.

Proof: Because of Lemma 1 there is an infinite sequence of points $z_1, z_2, z_3, \ldots, z_i \in E_d$ coverging to a limit point z_0 at which the real part of the function $h(z) = \ln f(z) = \ln c$ disappears. Since $f(z) \in K$, $h(z)$ is analytic in the neighborhood of E_d and $H(\theta) \equiv \mathrm{Re} h[\cosh(d + i\theta)]$ is a real analytic function of θ. $H(\theta)$ has an accumulation point of zeros at $\theta = \theta_0$ and therefore $H(\theta_0) = 0$. However the nth derivative of $H(\theta)$ having an accumulation point of zeros there also, vanishes at θ_0 and since $H(\theta)$ is real analytic it vanishes identically. Therefore $|f(z)| = c$ on the ellipse.

We now turn to the examination of the zero structure of $f(z) \in K$.

III. ZERO STRUCTURE

The functions $f(z) \in K$ have a finite number of zeros inside the ellipse E_d. The zeros may be classified as follows:

Class A: double zeros anywhere on the real axis, x_i, $i = 1, \ldots, L$;

Class B: complex conjugate zeros, a_i, a_i^*, $i = 1, \ldots, M$;

Class C: single real zeros for $1 < u_i < \cosh d$, $i = 1, \ldots, N$;

Class D: single real zeros for $-\cosh d < v_i < -1$, $i = 1, \ldots, P$.

An explicit form for $f \in K$ can be given in terms of its zeros when $f(z)$ is minimal by using the result of Lemma 2 that $|f(z)| = c$ on the ellipse. If $f(z) = c \exp h(z)$, then $h(z)$ has logarithmic singularities at the zeros of $f(z)$ and in additon the real part of $h(z)$ vanishes for $z \in E_d$. We can therefore write $f(z)$ as

$$f(z) = c \exp\left(2 \sum_{i=1}^{L} G(z, x_i) + \sum_{i=1}^{M} [G(z, a_i) + G(z, a_i^*)] + \sum_{i=1}^{N} G(z, u_i) + \sum_{i=1}^{P} G(z, v_i)\right), \tag{6}$$

where $G(z, z_i)$ is the Green function for the ellipse which has a logarithmic singularity at $z = z_i$ and has vanishing real part on the ellipse. $G(z, z_i)$ can be given in terms of Jacobi elliptic functions[6] and we can then write $f(z)$ when z is real ($z = x$) as

$$\begin{aligned} f(x) = c \prod_{n=1}^{L} & \left| k\,\mathrm{sn}[(K/\pi)(\theta + \theta_n)]\mathrm{sn}[(K/\pi)(\theta - \theta_n)]\right|^2 \\ & \times \prod_{n=1}^{M} \left| k\,\mathrm{sn}[(K/\pi)(\theta + \phi_n + id_n)]\,\mathrm{sn}[(K/\pi)(\theta - \phi_n + id_n)]\right|^2 \\ & \times \prod_{n=1}^{N} k\,\mathrm{sn}[(K/\pi)(\theta + ib_n)]\,\mathrm{sn}[K/\pi(\theta - ib_n)] \\ & \times \prod_{n=1}^{P} k\,\mathrm{sn}[(K/\pi)(\theta + \pi + ie_n)]\,\mathrm{sn}[(K/\pi)(\theta + \pi - ie_n)]. \end{aligned} \tag{7}$$

where $\cos\theta_n = x_n$, $\cos(\phi_n + id_n) = a_n$, $\cosh b_n = u_n$, $\cosh e_n = -v_n$, and $\cos\theta = x$. We have dropped the argument indicating the dependence of the $\mathrm{sn}(x)$ on the parameter k.[7] $K = K(k)$ is a complete elliptical integral and the parameter k is to be determined from the equation $2d/\pi = K(k')/K(k)$ where $k' = (1 - k^2)^{1/2}$. If d is small then $1 - k = O[\exp(-1/d)]$ and $K = O(1/d)$. We list some needed properties of $\mathrm{sn}(x)$ in the Appendix.

If $f(x)$ in Eq. (7) is to be of class K, then $f(1) = 1$. This gives a single constraint on the positions of the zeros x_i, a_i, u_i, and v_i. This can be written as

$$\begin{aligned} c^{-1} = \prod_{n=1}^{L} & k^2[\mathrm{sn}((K/\pi)\theta_n)]^4 \prod_{n=1}^{M} k^2 \left|\mathrm{sn}[(K/\pi)(\phi_n + id_n)]\right|^4 \\ & \times \prod_{n=1}^{N} k \left|\mathrm{sn}[(K/\pi)ib_n]\right|^2 \prod_{n=1}^{P} k \left|\mathrm{sn}[(K/\pi)(\pi + ie_n)]\right|^2. \end{aligned} \tag{8}$$

We now use Eqs. (7) and (8) and the properties (A1)—(A6) of $\mathrm{sn}(z)$ given in the Appendix to show that we can severely restrict the possible zeros f may have if it is to have minimal norm.

The general idea in the lemmas to follow is to show that zeros of $f(x)$ in one class can be replaced by zeros of another class and that the resulting function $g(x)$ satisfies $g(x) \leq f(x)$ for $-1 \leq x \leq +1$ and $f(1) = g(1) = 1$. Then since $\|g\| \leq \|f\|$ we only need consider function g with a more restricted set of zeros than those of f. We begin with Lemma 3.

Lemma 3: If $f(x) \in K$ and $\|f\|$ is minimal on K then $f(z)$ does not have zeros of class B (complex conjugate) inside E_d.

Proof: We replace the zeros of class B by zeros of class A. We introduce a zero on the real axis at $x = \cos\phi$ and defining $Y \equiv \mathrm{sn}[(K/\pi)\phi]$ we require that

$$Y^2 = |\mathrm{sn}[(K/\pi)(\theta_n + id_n)]|^2 \equiv |Z|^2. \tag{9}$$

This ensures that Eq. (8), the normalization condition, remains satisfied. That Eq. (9) is possible is easily seen from properties (A2)–(A6). If $|Z|^2 \leq 1$, then ϕ is pure real and, if $|Z|^2 > 1$, then $\phi = \pi + id_n'$.

To show that the norm of the new function is less than that of the original one, we must show that

$$|\mathrm{sn}[(K/\pi)(\theta+\phi_n)]\,\mathrm{sn}[(K/\pi)(\theta-\phi_n)]|^2 \leq |\mathrm{sn}[(K/\pi)(\theta+\theta_n+id_n)]\,\mathrm{sn}[(K/\pi)(\theta-\theta_n+id_n)]|^2. \tag{10}$$

Using the addition theorem (A1) we obtain

$$\left(\frac{X^2-|Z|^2}{1-k^2X^2|Z|^2}\right)^2 \leq \frac{(X^2-Z^2)(X^2-Z^{*2})}{(1-k^2X^2Z^2)(1-k^2X^2Z^{*2})} \tag{11}$$

where $X = \mathrm{sn}(K/\pi\theta)$ and $Y^2 = |Z|^2$ has been used. Rearrangement gives

$$2X^2(|Z|^2 - \mathrm{Re}Z^2)(1-k^2X^4)(1-k^2|Z|^4) \geq 0 \tag{12}$$

which is true because of (A2). The equality obtains in the region $0 \leq X \leq 1$ only at $X = 0$ as it should.

We now eliminate zeros of class C by proving Lemma 4.

Lemma 4: If $f \in K$ and $\|f\|$ is minimal on K, then f does not have zeros of class C inside E_d.

Proof: We proceed as in Lemma 3, where now we replace the single zero of class C by a double zero of class A. Defining $Y = \mathrm{sn}[(K/\pi)ib_n']$, where b_n' is real and consequently Y is imaginary, we require

$$k|\mathrm{sn}[(K/\pi)ib_n]|^2 = k^2|\mathrm{sn}[(K/\pi)ib_n']|^4 = k^2|Y|^4. \tag{13}$$

Properties (A2) and (A4) show that a b_n' can be found such that Eq. (13) is satisfied.

Now we must show that

$$k|\mathrm{sn}[(K/\pi)(\theta+ib_n)]|^2 \geq k^2|\mathrm{sn}[(K/\pi)(\theta+ib_n')]|^4. \tag{14}$$

Again using the addition theorem we find that we must have

$$\frac{X^2+k|Y|^4}{1+k^3X^2|Y|^4} \geq k\left(\frac{X^2+|Y|^2}{1+k^2X^2|Y|^2}\right)^2 \tag{15}$$

which upon rearrangement gives

$$X^2(1-k|Y|^2)^3(1+k|Y|^2)(1-kX^2) \geq 0. \tag{16}$$

This inequality holds in the region $0 \leq X \leq 1$ as can be seen from (A2) and again the equality only obtains at $X = 0$.

We cannot completely eliminate the zeros of class D, but we can show the following lemma.

Lemma 5: If $f \in K$ and $\|f\|$ is minimal on K, then $f(z)$ can have at most two single zeros of class D, one at $z = -1$ and one at $-\cosh d < z < -1$.

Proof: First we show that we can increase the separation of any two zeros of class D. Let $Y_j = \mathrm{sn}[(K/\pi)(\pi + ie_j)]$, $u_j = \cosh(e_j - i\pi)$, $j = 1,2,3,4$ (Y_j is real) where Y_1, Y_2 correspond to the initial zeros at u_1, u_2 and Y_3, Y_4 correspond to the final zeros at u_3, u_4. The contribution to $f(z)$ from u_1 and u_2 can then be written as

$$k^2|\mathrm{sn}[(K/\pi)(\theta+\pi+ie_1)]|^2\,|\mathrm{sn}[(K/\pi)(\theta+\pi+ie_2)]|^2 = k^2\frac{(Y_1^2-X^2)(Y_2^2-X^2)}{(1-k^2Y_1^2X^2)(1-k^2Y_2^2X^2)} \tag{17}$$

and we must show that

$$\frac{(Y_1^2-X^2)(Y_2^2-X^2)}{(1-k^2Y_1^2X^2)(1-k^2Y_2^2X^2)} \geq \frac{(Y_3^2-X^2)(Y_4^2-X^2)}{(1-k^2Y_3^2X^2)(1-k^2Y_4^2X^2)} \tag{18}$$

with the condition that $Y_1^2Y_2^2 = Y_3^2Y_4^2$ to ensure that Eq. (8) remains valid. Rearrangement of Eq. (18) shows that the inequality is satisfied if

$$Y_1^2Y_2^2 = Y_3^2Y_4^2, \quad Y_3^2 + Y_4^2 - Y_1^2 - Y_2^2 \geq 0. \tag{19}$$

It is easy to see using condition (A5) that the constraints of Eq. (19) allow u_3 and u_4 to have larger separation than u_1 and u_2.

Now picking any two zeros of class D we can separate them until either $e_3 = d$ (u_3 reaches the ellipse) or $e_4 = 0$ ($u_4 = z = -1$). In the first case the contribution to $f(z)$ from u_3 just becomes unity [condition (A2)] and we are left with a single zero at u_4. In the second case $f(z)$ has a single zero at u_3 and a zero at $u_4 = -1$. Continuing this procedure we can move all the zeros except possibly one to the ellipse or to the point -1. The multiple zero at $z = -1$ can be written as the product of a multiple double zero (class A) and possibly a single zero so we are left with possibly a single zero at $z = -1$ and a single zero at $-\cosh d < z < -1$.

We now define the functions of class K' as those functions g such that $g \in K(c/k^2)$ (i.e., $|g| \leq c/k^2$ on the ellipse) and such that g has only zeros of class A in E_d. We prove the following lemma.

Lemma 6: $\min_{f\in K(c)}\|f\| \geq \min_{g\in K'}\|g\|$.

Proof: We substitute the possible single zeros allowed by Lemma 5 by double zeros at $z = -1$. We have

$$k^2\frac{(Y^2-X^2)(1-X^2)}{(1-k^2X^2Y^2)(1-k^2X^2)} \geq k^2Y^2\left[\frac{1-X^2}{1-k^2X^2}\right]^2 \tag{20a}$$

$$k\frac{Y^2-X^2}{1-k^2X^2Y^2} \geq kY^2\left[\frac{1-X^2}{1-k^2X^2}\right]^2 \tag{20b}$$

$$k\frac{1-X^2}{1-k^2X^2} \geq k\left[\frac{1-X^2}{1-k^2X^2}\right]^2, \tag{20c}$$

where Eq. (20a) is for two single zeros at $z = -1$ and $z < -1$, Eq. (20b) is for a single zero at $z < -1$ and Eq. (20c) is for a single zero at $z = -1$. The inequalities are satisfied for $0 \leq X \leq 1$ and for $k^{-1} \geq Y^2 \geq 1$ (corresponding to a zero in the region $-\cosh d < z < -1$.) However on the the ellipse the double zero factors on the right-hand side of Eqs. (20) have maxima of Y^2, $k^{-1}Y^2$, and k^{-1} respectively. Because of the restrictions on Y^2 these maxima are all less than k^{-2} and therefore $|g(z)| \leq c/k^2$ on the ellipse and consequently $g \in K'$. The lemma is then proved and of course remains valid even if f has no single zeros since if $f \in K(c)$ then $f \in K(c/k^2)$.

We now need only consider functions $g \in K'$. Since the only zeros of g are double zeros along the real axis, the function $h(z) = [g(z)]^{1/2}$ is also an analytic function

inside E_d. In fact let us define a function h to be of class C if $h(z)$ is analytic inside the ellipse E_d, $h(1)=1$, $|h(z)| \leq c^{1/2}/k \equiv c'$, if $z \in E_d$ and $h(z)$ is real for z real inside E_d. Also we define the norm of h as

$$\|h\| = \int_{-1}^{1} dx[h(x)]^2. \tag{21}$$

From Lemma 6 we have immediately that $\min_{f\in K}\|f\| > \min_{h\in C}\|h\|$.

In the next section we shall exploit the analyticity of h to give a lower bound for $\|h\|$.

IV. LEGENDRE EXPANSION AND MINIMAL BOUND

A function $h \in C$ which the discussion of the last section has led us to consider has a series expansion in terms of Legendre polynomials, $P_l(z)$ which is uniformly convergent inside and on the ellipse E_d. In the next lemma we show that we need only consider those functions h which belong to the subclass of function C_L which have only the first $L+1$ Legendre coefficients, a_l, nonzero.

Lemma 7: $\lim_{L\to\infty} I(L) = I$ where $I(L) = \min_{h\in C_L}\|h\|$ and $I = \min_{h\in C}\|h\|$.

Proof: It is clear that $\lim_{L\to\infty} I(L) \geq I$ since $C_L \subset C$.

Let $h(z) \in C$ be a function for which $\|h\| = I$. Denote the sum of its first $L+1$ Legendre terms by $h_L(z)$. If $\max_{z\in E_d} |h_L(z)/h_L(1)| \leq c'$, $h_L(z)/h_L(1) \in C_L$ and $\|h_L\| \geq I(L)[h_L(1)]^2$. If $\max_{z\in E_d} |h_L(z)/h_L(1)| \geq c'$ then we can find an $a>0$ such that

$$\max_{z\in E_d}\left|\frac{h_L(z)+a}{h_L(1)+a}\right| = c' \quad \text{and} \quad \frac{h_L(z)+a}{h_L(1)+a} \in C_L$$

so that $\|h_L(z)+a\| \geq I(L)[h_L(1)+a]^2$. Because of the uniform convergence of h_L, however, $h_L(1)-1$ and , $\max_{z\in E_d}|h_L(z)| - c'$ are arbitrarily small and consequently a is arbitrarily small for L large enough. Using the fact that $\|h\| \geq \|h_L\|$ we therefore see that $\|h\| = I \geq \lim_{L\to\infty} I(L)$ and therefore $I = \lim_{L\to\infty} I(L)$.

We shall now find a lower bound for $I(L)$. Any $h_L \in C_L$ can be expanded in the following two equivalent forms,

$$h_L(z) = \sum_{l=0}^{L} a_l(2l+1)P_l(z) = \sum_{n=0}^{L} A_n \cos n\theta \tag{22}$$

where $z = \cos\theta$ and θ is in general complex. The coefficients a_l and A_n are related by the equation[8]

$$a_l = \sum_{n=l}^{L} h_{nl}A_n \tag{23}$$

where

$$h_{nl} = \begin{cases} -\dfrac{n}{8}\dfrac{\Gamma((n-l-1)/2)\Gamma((n+l)/2)}{\Gamma((n-l+2)/2)\Gamma((n+l+3)/2)} & \\ \qquad \text{if } n \geq l, n-l \text{ even}, \ n \neq 0, & \\ 1 \quad \text{if } n=l=0, & \\ 0 \quad \text{otherwise.} & \end{cases} \tag{24}$$

To show that large values of l in Eq. (22) give a small contribution to $I(L)$, we will introduce a cutoff function and estimate the difference between the cutoff Legendre expansion and Eq. (22). We choose

$$c_l = \begin{cases} 1 \quad \text{if } l \leq K, & \\ 1-(K-l)^2 d^2/2 \quad \text{if } K \leq l \leq K+(1/d), & \\ \left(K-l+\dfrac{2}{d}\right)^2 d^2/2 \quad \text{if } K+(1/d) \leq l \leq K+(2/d), & \\ 0 \quad \text{if } l \geq K+(2/d), & \end{cases} \tag{25}$$

where K is a parameter to be determined later. It is easy to see that c_l satisfies the following inequalities,

$$\begin{aligned} &0 \leq c_{l-2} - c_l \leq 2d, \\ &|c_l + c_{l-4} - 2c_{l-2}| \leq 4d^2. \end{aligned} \tag{26}$$

The difference between the cutoff Legendre expansion and Eq. (22) at $\theta = 0$ is defined by

$$\Delta = \sum_{l=0}^{L} a_l(2l+1)(1-c_l) \tag{27}$$

and is given by

$$\begin{aligned} \Delta &= \sum_{n=1}^{L}{}' \sum_{l=K}^{L} h_{nl}A_n(2l+1)(1-c_l) \\ &= \sum_{n=K}^{L} B_n \sum_{l=K}^{n}{}' \frac{\Gamma((n-l-1)/2)}{\Gamma((n-l+2)/2)}[s_{n,l} - s_{n-2,l-2}] \\ &= \sum_{n=K}^{L} B_n \sum_{l=K}^{n}{}' \frac{\Gamma((n-l-1)/2)}{\Gamma((n-l+2)/2)}[s_{n,n} - s_{n-2,n-2} \\ &\quad - (n-l)(s_{n,n} - s_{n-2,n-2} - s_{n,l} + s_{n-2,l-2})/(n-l)], \end{aligned} \tag{28}$$

where $B_n = \sum_{m=n}^{L}{}' A_m$. The prime on the sums means that the summation is only over even values of $n-l$ or $n-m$. Also we define

$$s_{n,l} = -\frac{n}{8}\frac{\Gamma((n+l)/2)}{\Gamma((n+l+3)/2)}(2l+1)(1-c_l). \tag{29}$$

Using the elementary relation

$$\sum_{k=0}^{m}\frac{\Gamma(k-s)}{k!} = -\frac{1}{s}\frac{\Gamma(m+1-s)}{m!} \tag{30}$$

and the inequality

$$\left|\frac{s_{n,n} - s_{n-2,n-2} - s_{n,l} + s_{n-2,l-2}}{(n-l)}\right| \leq g_n \tag{31}$$

where

$$g_n = \max_{l;\,0\leq l\leq n}\left|s_{n,l} - s_{n-2,l-2} - s_{n,l-2} + s_{n-2,l-4}\right|/2 \tag{32}$$

we obtain the estimate

$$\begin{aligned} |\Delta| &\leq \sum_{n=K}^{\infty}|B_n|\left\{4\left[\frac{n}{2}\right]g_n\frac{\Gamma([n/2]+\frac{1}{2})}{[n/2]!}\right. \\ &\quad \left. + 2|s_{n,n} - s_{n-2,n-2}|\frac{\Gamma([(n-K)/2]+\frac{1}{2})}{[(n-K)/2]!}\right\}, \end{aligned} \tag{33}$$

where $[\nu]$ denotes the integer part of ν.

We now estimate each term of Eq. (33) separately. Using Eqs. (25) and (29) we easily find that

$$2|s_{n,n} - s_{n-2,n-2}| \leq \frac{\Gamma(n-1)}{\Gamma(n-\frac{1}{2})}(1+2nd). \tag{34}$$

A more lengthy but straightforward calculation gives

$$g_n \leq \frac{\Gamma(n/2-3)}{\Gamma((n-3)/2)}(2+3dn+\tfrac{1}{2}d^2n^2). \tag{35}$$

Finally we give a bound on B_n by writing

$$B_n = \sum_{m=n}^{L}{}' A_m = (1/\pi)\int_0^{2\pi} \sum_{m=0}^{L} A_m \cos[m(\theta + id)] \sum_{k=n}^{L}{}' \exp[-k(d - i\theta)]\, d\theta. \tag{36}$$

Using Eq. (22) and the bound on $h_L(z)$ on the ellipse we have

$$|B_n| \leq \frac{c'}{\pi}\int_0^{2\pi} \left| \frac{\exp[-n(d+i\theta)] - \exp[-L(d+i\theta)]}{1 - \exp[-2(d+i\theta)]} \right| d\theta. \tag{37}$$

For L large enough the second term in the numerator can be ignored and we can write

$$|B_n| \leq \pi^{-1} 2^{1/2} c' \exp[-(n-1)d] \times \int_0^{\pi} (\cosh 2d - \cos\theta)^{-1/2}\, d\theta = c^{1/2}\exp(-nd)h(d), \tag{38}$$

where

$$h(d) = 2\pi^{-1}[\ln(1/d) + O(1)] \tag{39}$$

as $d \to 0$. Recall that $c' = c^{1/2}k^{-1}$ and $k = 1 - \exp(-1/d)$ for small d.

We obtain a bound on Δ by substituting Eqs. (34), (35), (38), and (39) into Eq. (33). This yields

$$|\Delta| \leq 4c^{1/2}h(d)\exp(-Kd)[(Kd) + 3(Kd)^{1/2} + 8 + 3(Kd)^{-1/2} + 12(Kd)^{-1}] \tag{40}$$

where we have made the approximations that $K \gg 1$ and $d \ll 1$.

We now choose K such that

$$|\Delta| \leq \{\ln[c^{1/2}h(d)]\}^{-1}. \tag{41}$$

This can easily be done by taking

$$Kd = \ln\{\lambda(\ln\lambda\rho)^2[1 + 3(\ln\lambda)^{-1/2}]\} \tag{42}$$

where $\lambda = 4c^{1/2}h(d)$ and where ρ satisfies the inequality

$$\rho \geq \exp(8 + 3c^{-1/4} + 12c^{-1/2})[\ln(\lambda\rho)]^2[1 + 3(\ln\lambda)^{-1/2}]. \tag{43}$$

Using the definition of Δ in Eq. (27) and the bound of Eq. (41) and the fact that $h_L(1) = 1$, we obtain

$$\sum_{l=0}^{L} (2l+1)a_l c_l = 1 + \eta\{\ln[c^{1/2}h(d)]\}^{-1}, \tag{44}$$

where $-1 \leq \eta \leq 1$. Equation (44) is a constraint on the expansion coefficients a_l.

We are now able to place a lower bound on $I(L)$. Because of the orthogonality of the Legendre polynomials we can write

$$I(L) = \sum_{l=0}^{L} 2(2l+1)a_l^2. \tag{45}$$

The original problem was to minimize Eq. (45) subject to the condition that $|h_L(z)| = c^{1/2}$ on the ellipse. We certainly obtain a lower bound, however, if we ignore the constraint on the ellipse and use only Eq. (44). Using Lagrange multipliers we find immediately that the minimum of $I(L)$ subject to the constraint Eq. (44) occurs at $a_l = \alpha c_l$ where from Eq. (44)

$$\alpha = (1 + \eta\{\ln[c^{1/2}h(d)]\}^{-1})[\textstyle\sum(2l+1)c_l^2]^{-1}. \tag{46}$$

We then obtain

$$I(L) = 2[1 + \eta\{\ln[c^{1/2}h(d)]\}^{-1}]^2[\textstyle\sum(2l+1)c_l^2]^{-1}. \tag{47}$$

The sum can be estimated using Eq. (25) and finally we arrive at the bound

$$I(L) \geq 2\frac{(1 - \{\ln[c^{1/2}h(d)]\}^{-1})^2}{(K+1)^2 + 2(2K + 2/d + 1)/d}. \tag{48}$$

This bound is valid for nonasymptotic values of c as well. If $c \to \infty$ and $1/d \to \infty$ such that $\ln(c)/\ln(1/d)$ remains finite as in the case of scattering processes where c and $1/d$ are proportional to a power of s, the leading term in Eq. (48) is, using Eqs. (39) and (42),

$$I(L) \geq 2\,d^2(\ln c^{1/2})^{-2}. \tag{49}$$

Keeping the first nonleading term in Eq. (42), we obtain

$$I(L) \geq 2d^2\{\ln[c^{1/2}\ln(1/d)(\ln c^{1/2})^2]\}^{-2}. \tag{50}$$

This provides a proof of the theorem stated in the Introduction.

V. CONCLUSION

The theorem we have proved will be used in applications to give rigorous bounds on inclusive cross sections.

We wish to indicate in conclusion that the bound of Eq. (50) cannot be substantially improved. If we ignore the constraint condition on the ellipse and merely minimize $I(L)$ subject to the normalization condition $h_L(1) = 1$ we find trivially that

$$I(L) = 2(L+1)^{-2} \tag{51}$$

with the Legendre coefficients given by $a_l = (L+1)^{-2}$. The The magnitude of $h_L(z)$ on the ellipse is then

$$|h_L(z)| = |(L+1)^{-2} \times \sum_{l=0}^{L}(2l+1)P_l(z)| = \left|\frac{P_{L+1}(z) - P_L(z)}{(L+1)(z-1)}\right|. \tag{52}$$

Using the asymptotic form for the Legendre polynomials[9]

$$P_L(z) \sim (2\pi L)^{-1/2}(z^2-1)^{-1/4}[z + (z^2-1)^{1/2}]^{L+1/2} \tag{53}$$

we see that if we require that

$$\exp(dL)(dL)^{-3/2} \leq c^{1/2} \tag{54}$$

we then have $|h_L(z)| \leq c^{1/2}$ on the ellipse. Solving Eq. (53) for L we obtain

$$L = d^{-1}\ln[c^{1/2}(\ln c^{1/2})^{3/2}] \tag{55}$$

which gives for Eq. (51)

$$I(L) \approx 2\,d^2\{\ln[c^{1/2}(\ln c^{1/2})^{3/2}]\}^{-2}. \tag{56}$$

Therefore we have found an explicit function with an $I(L)$, Eq. (56), which differs from our general result only in nonleading asymptotic terms.

APPENDIX

We list here some properties of the Jacobi elliptic sine functions.[7]

(A1) Addition theorem

$$\mathrm{sn}(u+v)\,\mathrm{sn}(u-v)=\frac{(\mathrm{sn}u)^2-(\mathrm{sn}v)^2}{1-k^2(\mathrm{sn}u\,\mathrm{sn}v)^2}.$$

(A2) $k\,|\,\mathrm{sn}[(K/\pi)(\theta+i\,d_n)]\,|^2\leq 1$. The equality sign only holds on the ellipse where $d_n=d$. For $d_n=0$, $k[\mathrm{sn}((K/\pi)\theta)]^2\leq k<1$.

(A3) $\mathrm{sn}(0)=0$, $\mathrm{sn}(K)=1$, $|\,\mathrm{sn}[(K/\pi)(\theta+i\,d)]\,|=k^{-1/2}$.

(A4) $|\,\mathrm{sn}((K/\pi)\,i\,d_n)\,|^2=-[\mathrm{sn}((K/\pi)i\,d_n)]^2$ is a monotonically increasing function of d_n for $0\leq d_n\leq d$.

(A5) $|\,\mathrm{sn}[(K/\pi)(\pi+i\,d_n)]\,|^2=\{\mathrm{sn}[(K/\pi)(\pi+i\,d_n)]\}^2$ is a monotonically increasing function of d_n for $0\leq d_n\leq d$.

(A6) $\mathrm{sn}((K/\pi)\theta)$ is a monotonically increasing function of θ for $0\leq\theta\leq\pi$.

*Research supported in part by the National Science Foundation under Grant No. GP-43672.

[1]G. Tiktopoulos and S.B. Treiman, Phys. Rev. **167**, 1408 (1968).
[2]G. Tiktopoulos and S.B. Treiman, Phys. Rev. D **6**, 2045 (1972).
[3]J.J.G. Scanio and P. Suranyi, Phys. Rev. D **10**, 2954 (1974).
[4]A.A. Logunov and M.A. Mestvirishvily, CERN Report No. TH-1707, (1973) (unpublished) and references therein.
[5]For a reveiw of rigorous bounds on two particle cross sections see R.J. Eden, Rev. Mod. Phys. **43**, 15 (1971).
[6]The Green function is given in infinite series form in P. Morse and H. Feshbach, *Methods of Theoretical Physics* (McGraw-Hill, New York, 1953), Vol. 2, p. 1202, and the Jacobi form can be derived from this. One can obtain an equivalent representation of $f(x)$ by noting that the conformal map from the ellipse to the circle is given by $w=k^{1/2}\,\mathrm{sn}(2K/\pi\arcsin z)$.
[7]We use the notation of E.T. Whittaker and G.N. Watson, *Modern Analysis* (Cambridge U.P., Cambridge, 1965), 4th ed.
[8]I.S. Gradshteyn and I.M. Ryzhik, *Table of Integrals Series and Products* (Academic, New York, 1965), p. 824, 7.245,2.
[9]See, for example, G. Szegö, *Orthogonal Polynomials* (American Mathematical Society, New York, 1959), p. 188.

Extension of the statistical mechanics of equilibrium to noncommutative constraints

Elihu Lubkin

Department of Physics, University of Wisconsin-Milwaukee, Milwaukee, Wisconsin 53201
(Received 11 December 1975)

A formula similar to the Gibbs canonical and grand canonical ensembles is proven for the ensemble of maximal entropy among those ensembles of common mean values of possibly *noncommuting* operators. This is done over a Hilbert space of finite dimension Δ. A partition matrix Π becomes important; the partition function $Z = \mathrm{Tr}\Pi$ displaces Π in the thermodynamics only in the commutative case. Generalization to Δ infinite is discussed informally.

Maximizing the entropy subject to an ensemble-mean energy constraint produces the Gibbs canonical ensemble. If also a mean particle number is imposed, one gets the Gibbs grand canonical ensemble. The energy and particle-number operators H and N are usually taken to commute; also one can go further to several commuting particle-number operators. These familiar examples are solved by a familiar Boltzmann-factor formula, (4) below. Von Neumann[1,2] shows that the desired ensemble P must commute with H in the "canonical" problem by using measurement theory: A noncommuting ensemble would be converted to an energy-diagonal ensemble of greater entropy by measurement of a degeneracy-lifting energy-commuting quantity "R"; this proof also easily establishes the similar theorem for "grand canonical" cases:

Von Neumann's *theorem*: The ensemble P which maximizes entropy subject to several simultaneous constraints stated in terms of commuting operators must itself commute with them all.

One then chooses a representation with everything diagonal, and does the usual Lagrange-multiplier calculation.

The point of the present paper is that (4) is valid also for noncommuting constraint operators, even though the calculation which establishes its proof must proceed without the luxury of a prior diagonalization.

What naturally interesting problem presents us with an example? A system which freely exchanges not only energy and particle number(s) with a surround, but also exchanges x, y, and z components J_1, J_2, J_3 of angular momentum comes to mind; of course, these do not commute with each other. In this case an appeal to symmetry allows one to, however, avoid the noncommutative problem (see #10 below). Indeed, I have not yet found a natural application wherein knowing (4) for the noncommutative case cannot be circumvented. Nevertheless, this is the general solution:

#1. *Lemma*: The ensemble P over a state Hilbert space of finite dimension Δ, which maximizes the entropy

$$S = -\mathrm{Tr}P\ \ln P, \tag{1}$$

subject to $n+1$ constraints of form

$$\mathrm{Tr}PA_j = Q_j, \quad j = 0, 1, \ldots, n, \tag{2}$$

where $A_0 = I$, the unit matrix, and $Q_0 = 1$ conveniently express normalization, where the A_j are Hermitian matrices,[3] and the Q_j are constants, is the generalized Gibbs ensemble

$$P = \Pi/Z, \quad Z = \mathrm{Tr}\Pi, \tag{3}$$

$$\Pi = \exp(-\sum_{j=1}^{n} \lambda_j A_j). \tag{4}$$

As the "Lagrange multipliers" λ_j range freely over the reals ($j = 1, \ldots, n$), the $Q_1, \ldots, Q_n$ take on all possible values for which a solution exists.[4]

#2. *Comment*: Though #1 itself is the point of this paper, I call it a "lemma" in order to convey a sense of incompleteness, and this for two reasons. One is the lack of any dramatic application. The other is the restriction to finite Δ. This restriction is largely responsible for the freedom of the λ_j, including for example negative temperature; see #11 below.

#3. *Proof* of #1: P is expressed in terms of its eigenvalues p_a assembled as a diagonal matrix ρ, and a unitary transformation $U = X + iY$, X and Y real. The real variables p_a, x_{ab}, y_{ab} (elements of X and Y) are to be the independent variables in the method of Lagrange multipliers,

$$P = U\rho U^{-1}; \tag{5}$$

$S = -\mathrm{Tr}P\ \ln P = -\mathrm{Tr}U\ \rho \ln\rho\ U^{-1} = -\mathrm{Tr}\rho\ln\rho$ is seen not to explicitly involve U. The unitarity of U will be expressed by treating $(X+iY)({}^tX - i\,{}^tY) = I$ as $2\Delta^2$ real red-tape constraints, along with the $n+1$ physical constraints $\mathrm{Tr}PA_j = Q_j$. The elements of the A_j matrices will also be expressed in terms of reals, $A_{jbc} = B_{jbc} + iC_{jbc}$. Hence the explicit real-variable calculation reads as follows: Maximize

$$S = -\sum_i p_i \ln p_i$$

subject to the constraints

$$Q_j = \sum_{abc} p_a(x_{ba} - iy_{ba})(B_{jbc} + iC_{jbc})(x_{ca} + iy_{ca}) \tag{6}$$

and

$$\delta_{ac} = \sum_b (x_{ab} + iy_{ab})(x_{cb} - iy_{cb}). \tag{7}$$

dS is equated to a real linear combination of the real parts of the dQ_j (coefficients λ_{jR}), of the imaginary

parts of the dQ_j (coefficients λ_{jI}), and of the real and imaginary parts of the $d\delta_{ac}$ (coefficients λ_{acR} and λ_{acI}, respectively).[5] Then the dp_i terms, the dx_{pq} terms, and the dy_{pq} terms must balance separately, producing the following equations.

The dp_i equations:

$$-1-\ln p_i=\sum_{ac}\lambda_{acR}\cdot 0+\sum_{ac}\lambda_{acI}\cdot 0$$
$$+\sum_j\lambda_{jR}\sum_{bc}(x_{bi}B_{jbc}x_{ci}-x_{bi}C_{jbc}y_{ci}$$
$$+y_{bi}B_{jbc}y_{ci}+y_{bi}C_{jbc}x_{ci})$$
$$+\sum_j\lambda_{jI}\sum_{bc}(y_{bi}C_{jbc}y_{ci}-y_{bi}B_{jbc}x_{ci}$$
$$+x_{bi}C_{jbc}x_{ci}+x_{bi}B_{jbc}y_{ci}). \tag{8}$$

The dx_{pq} equations:

$$0=\sum_{ac}\lambda_{acR}(\delta_{pa}x_{cq}+\delta_{pc}x_{aq})+\lambda_{acI}(-\delta_{pa}y_{cq}+\delta_{pc}y_{aq})$$
$$+\sum_j\lambda_{jR}p_q\sum_{bc}(\delta_{pb}B_{jbc}x_{cq}-\delta_{pb}C_{jbc}y_{cq}+$$
$$+x_{bq}B_{jbc}\delta_{pc}+y_{bq}C_{jbc}\delta_{pc})$$
$$+\sum_j\lambda_{jI}p_q\sum_{bc}(\delta_{pb}C_{jbc}x_{cq}+\delta_{pb}B_{jbc}y_{cq}$$
$$-y_{bq}B_{jbc}\delta_{pc}+x_{bq}C_{jbc}\delta_{pc}). \tag{9}$$

The dy_{pq} equations:

$$0=\sum_{ac}\lambda_{acR}(\delta_{pa}y_{cq}+\delta_{pc}y_{aq})+\lambda_{acI}(\delta_{pa}x_{cq}-\delta_{pc}x_{aq})$$
$$+\sum_j\lambda_{jR}p_q\sum_{bc}(\delta_{pb}B_{jbc}y_{cq}+\delta_{pb}C_{jbc}x_{cq}$$
$$-x_{bq}C_{jbc}\delta_{pc}+y_{bq}B_{jbc}\delta_{pc})$$
$$+\sum_j\lambda_{jI}p_q\sum_{bc}(\delta_{pb}C_{jbc}y_{cq}-\delta_{pb}B_{jbc}x_{cq}$$
$$+y_{bq}C_{jbc}\delta_{pc}+x_{bq}B_{jbc}\delta_{pc}). \tag{10}$$

Now define complex quantities

$$\lambda_j=\lambda_{jR}+i\lambda_{jI},\quad \lambda_{ac}=\lambda_{acR}+i\lambda_{acI}; \tag{11}$$

it is also convenient to define a matrix A by

$$\sum_j A_{jbc}\lambda_j^*=A_{bc}. \tag{12}$$

Then (9) plus i times (10) abbreviates to

$$0=(\Lambda+\Lambda^\dagger)U+(A+A^\dagger)U\rho,$$

where Λ is the matrix of λ_{ac}'s. Right multiplication by U^{-1} produces

$$0=\Lambda+\Lambda^\dagger+(A+A^\dagger)P. \tag{13}$$

The sense of (13) is that $(A+A^\dagger)P$ is Hermitian. Since $A+A^\dagger$ and P are both Hermitian, this is equivalent to commutativity of P with $A+A^\dagger$. That is all that we have involving the red-tape constraints' multipliers.

Equation (8) in turn condenses to

$$-1-\ln p_i=(U^\dagger\tfrac{1}{2}(A+A^\dagger)U)_{ii}; \tag{14}$$

in words, $-1-\ln\rho$ is the diagonal of $U^\dagger\frac{1}{2}(A+A^\dagger)U$. But commutativity of P with $A+A^\dagger$ is equivalent to commutativity of ρ with $U^\dagger\frac{1}{2}(A+A^\dagger)U$; hence we may choose a representation wherein both ρ and $U^\dagger\frac{1}{2}(A+A^\dagger)U$ are simultaneously diagonal. Indeed, the original representation already has ρ diagonal, and if the other matrix is not already diagonal in consequence, that is only because diagonality of ρ does not entirely specify choice of representation: Our new choice at this point may be regarded as having been made at the outset, and hence does not undo anything proven. Equation (14) is thus improved to

$$-1-\ln\rho=U^\dagger\tfrac{1}{2}(A+A^\dagger)U,$$

or

$$-1-\ln P=\tfrac{1}{2}(A+A^\dagger)=\sum_j A_j\,\mathrm{Re}\lambda_j. \tag{15}$$

It is only in the last step that the assumption that the A_j are Hermitian was introduced. The formula (4) now follows upon writing $\mathrm{Re}\lambda_j$ as λ_j, now considered real.

The only conditions required for a proper ensemble P not yet stated are nonnegativity of the eigenvalues p_i, but that is automatic from (4), since an exponential of an Hermitian matrix is necessarily positive. Then any n reals $\lambda_1,\ldots,\lambda_n$ will, through (4,3,2), generate some list of $Q_1,\ldots,Q_n$. Finally, for any such achievable list of $Q_1,\ldots,Q_n$ values, there is only one local stationary entropy ensemble because the infragraph of the entropy as a function on the real space of ensembles is a convex body. There is also one maximum[6]; hence the Lagrange-multiplier search for a local stationary point automatically discovers the unique maximum, except when the set of ensembles which meet the constraints is empty. QED

#4. *f-Entropy*: If, instead of using the function $x\to -x\ln x$ in defining the entropy, one uses $S=\mathrm{Tr}[f(P)]$, f convex upwards, $-1-\ln x$ is replaced by f', one keeps $j=0$ together with $j=1,\ldots,n$, and ϕ such that $\phi(f'(x))=x$ replaces the exponential operation $x\to\exp[-(1+x)]$. Nonnegativity is no longer automatic.

#5. *Thermodynamics*: In the Lagrange multiplier method, dS for S maximal is expressed as a linear combination of constraints,

$$dS=\sum_{j=0}^{n}\lambda_{jR}\,d\,\mathrm{Re}Q_j+\lambda_{jI}\,d\,\mathrm{Im}Q_j+\text{red-tape terms}. \tag{16}$$

The red-tape and $j=0$ terms vanish if we impose unitarity and normalization. Also, the use of A_j Hermitian requires Q_j to be real for a nonvacuous case. Hence (16) degenerates to

$$dS=\sum_{j=1}^{n}\lambda_j\,dQ_j,$$

with the λ_j and Q_j real. Thus, $\lambda_j=\partial S/\partial Q_j$ at fixed Q_k, $k\neq j$. This may be called a "generalized thermal vector," in recognition of the fact that when the Q_j are additively conserved quantities, it is the λ_j which must balance between two systems in order that a flow will not increase the entropy: There is a deduction of a "zeroth law of thermodynamics" from the second law with respect to each λ_j belonging to an additively conserved Q_j. For the Q list: energy, particle number, volume, angular momentum component, linear momentum component, the corresponding λ are $1/T$, $-\mu/T$, P/T, $-\omega/T$, $-v/$

T, where T, μ, P, ω, v, are absolute temperature (in energy units), chemical potential (with Avogadro's number set equal to 1), pressure, angular velocity component, linear velocity component, as usual.

Expression of the entropy as a function of the generalized thermal parameters proceeds as usual: (1) and (3) yield

$$S = \ln Z - Z^{-1}\,\mathrm{Tr}\Pi \ln\Pi,$$

then (4) gives

$$S = \ln Z - \sum_{j=1}^{n} \lambda_j Q_j, \tag{17}$$

where

$$Q_j = Z^{-1}\,\mathrm{Tr}\Pi A_j. \tag{18}$$

The sequence (4,3,18,17) explicitly parametrizes everything in terms of the generalized thermal parameters λ_j. These formulas are conventional. Noncommutativity only makes it impossible to further simplify (18) to

$$Q_j = -\frac{\partial \ln Z}{\partial \lambda_j}, \quad \text{(false)}, \tag{19}$$

although (19) is "true along a ray" in the following sense: If $\lambda_j = \beta\alpha_j$ and $\partial/\partial\beta$ is defined with the α_j fixed, then

$$\sum_{j=1}^{n} \lambda_j Q_j = -\frac{\partial \ln Z}{\partial \ln \beta}. \tag{20}$$

Thus, one needs to know the "partition matrix" Π, not only its trace Z ("partition function"), in order to reduce a noncommutative statistical mechanics to a thermodynamics.

In spite of the absence of any *dramatic* example, I give a few immediate consequences of the Lemma.

#6. *Von Neumann's theorem*: This is a corollary by inspection of (4) in the case that the A_j are given to commute, P being explicitly a function of them.

#7. *Invariance of P under a symmetry group*: If a group of unitary transformations $\{U_g\}$, $g \in G$, commutes with each A_j, then evidently $U_g P U_g^{-1} = P$ for all g. This illustrates Ref. 6.

#8. *Thermodynamic axis*: In the three-component angular momentum example, J_1, J_2, J_3 appear in P only in the combination $\lambda_1 J_1 + \lambda_2 J_2 + \lambda_3 J_3 = -(1/T)\boldsymbol{\omega}\cdot\mathbf{J}$. Hence the equilibrium ensemble depends only on the angular momentum operator along a single axis, indeed, along the axis parallel to the angular velocity $\boldsymbol{\omega}$ externally imposed by the angular velocity reservoir.

#9. *Are commutators too small to matter?*: Is noncommutativity unimportant in thermodynamics, because commutators are of order $\hbar$? First, even if this is so in an example, the result (4) is useful for checking it. One would probably wish to show that dropping commutators by using (19) instead of (18) is a good approximation.

Second, commutators should be important when both noncommutativity and a microscopic system are prominent. Recall that kinetic theory is Maxwell's thermodynamics of a *single molecule* modified by "collisions."

#10. *Angular momentum without the lemma*: Replace $\mathrm{Tr}PJ_k = Q_k$, $k = 1,2,3$ by their linear combinations

$$\mathrm{Tr}PJ_1' = 0, \quad \mathrm{Tr}PJ_2' = 0, \tag{21}$$

$$\mathrm{Tr}PJ_3' = (Q_1^2 + Q_2^2 + Q_3^2)^{1/2} \tag{22}$$

obtained through a rotation of axes. Then *drop* Eqs. (21), and solve the now commutative "associated" problem (which may involve other operators, like the energy H, which commute with J_3'). The solution P of this associated problem will be found to satisfy (21) because the conditions of the associated problem are invariant to the one-dimensional subgroup of rotations about the $3'$ axis.[6] (Finiteness of Δ is being employed.) *A fortiori* P satisfies the original problem in which (21) are also imposed: A maximum over some domain remains maximum over a restricted domain, from which it is not itself excluded.

Of course, deft avoidance of commutators may also be practiced in conjunction with (4) and (20).

#11. $\Delta \to \infty$?: In von Neumann's book[2] we are reassured that $\mathrm{Tr}AB = \mathrm{Tr}(A^{1/2}B^{1/2})^{\dagger}A^{1/2}B^{1/2}$ will appear only for nonnegative operators A and B, and that difficulties attending conditional convergence will therefore be moot. The only infinite phenomena remaining are convergence of definite trace sums to $+\infty$. This assurance is not available for $\mathrm{Tr}PN$, if N is a particle number allowed to be negative because of antiparticles, or for $\mathrm{Tr}PJ_3$. The notion of limit as Δ goes infinite must be shaped by detailed physical considerations, if the mathematical possibilities are thus opened to a relatively undefined variety of limits. Furthermore, nothing is proven here for Δ infinite; a discussion based on formulas (1)–(4) requires a prelimit context wherein Δ is finite. As long as Δ is finite, there is no impossible "thermal vector" $(\lambda_1, \ldots, \lambda_n)$. An indication of how some thermal vectors become impossible is to study the convergence of the normalization $Z = \mathrm{Tr}\exp(-\sum_{j=1}^{n} \times \lambda_j A_j)$ along a ray, $\lambda_j = \beta\alpha_j$, with β variable, the α_j fixed. The operator is like the canonical ensemble operator $\exp(-\beta H)$, with $\sum_{j=1}^{n} \alpha_j A_j \equiv H(\alpha)$ in place of H. If, as $\Delta \to \infty$, the spectrum of $H(\alpha)$ becomes unbounded above (below), β may not be negative (positive); a sufficiently dense concentration of $H(\alpha)$ "level density" will restrict β even further.

[1]Anthony Lomazzo stressed the importance of such a discussion in rendering the usual textbook treatment cogent, in a private conversation.

[2]John von Neumann, *Mathematical Foundations of Quantum Mechanics*, translated by R. T. Beyer (Princeton U. P., Princeton, N.J., 1955).

[3]The situation for any non-Hermitian matrices A_j is easily obtained by breaking such A_j into Hermitian and skew-Hermitian parts. Hence the restriction to Hermitian A_j is to be regarded as an inessential convenience.

[4]No solution will exist for example in cases where equations (2) are obviously incompatible, e.g., $A_1 = A_2$ but $Q_1 \neq Q_2$.

[5]Indulgence in the common abuse of language wherein a symbol designating a function's value is impressed to do more is noted.

[6]Elihu Lubkin, J. Math. Phys. **16**, 837 (1975).

Solitons and simple pseudopotentials

James Corones

Department of Mathematics, Iowa State University, Ames, Iowa 50011
(Received 20 October 1975)

A simple type of pseudopotential is defined. It is shown how to compute this pseudopotential by classical means. When the construction is attempted based on the Hirota equation it is found that the condition necessary for the existence of the pseudopotential is the same as that for the existence of n-soliton solutions to the Hirota equation. It is also shown how to obtain Backlund transformations using this pseudopotential. It is also pointed out that without making any closure assumption this pseudopotential defines a Lie algebra.

I. INTRODUCTION

Wahlquist and Estabrook recently introduced the concept of a pseudopotential into the study of nonlinear partial differential equations.[1] It is the purpose of this paper to show by example that one simple type of pseudopotential is of particular importance in the study of equations having soliton solutions and further to show how to exhaustively and directly compute these pseudopotentials by purely classical means, as opposed to using the language of differential forms employed by Wahlquist and Estabrook.

To accomplish this the relevant definitions are given: those of a pseudopotential, a pseudopotential of the first kind, and a non-Abelian pseudopotential of the first kind. With these linguistic matters attended to, a derivation of of a necessary condition for the Hirota equation[2] to have a non-Abelian pseudopotential of the first kind is given. This condition is the same as that given by Hirota for his equation to have an n-soliton solution.[2] Next the Hirota equation is specialized to the nonlinear Schrödinger equation, and an expression is derived that gives a solution of the nonlinear Schrödinger equation in terms of another, distinct solution and the pseudopotential of the first kind associated with the nonlinear Schrödinger equation. From this Lamb's[3] Bäcklund transformation for the nonlinear Schrödinger equation can be recovered. Finally, evidence is presented in support of the conjecture that the existence of a non-Abelian pseudopotential of the first kind associated with a given evolution equation is a necessary condition for the existence of soliton solutions to the given equation. The importance of affirming this conjecture is discussed.

II. SOME DEFINITIONS

Consideration will be restricted to equations of the form

$$\varphi_t = K(\varphi, \bar{\varphi}, \varphi_x, \bar{\varphi}_x, \cdots), \tag{1}$$

where K is some function of $\varphi, \bar{\varphi}$ and their spatial derivatives up to order $m+1$. Let S be the set of $\varphi, \bar{\varphi}$ and all of their spatial derivatives up to order m. The set of all *pseudopotentials* associated with (1) is the set of all functions $q^i(x,t)$, $i=1,\ldots,n$, n arbitrary, such that

$$q^j_x = A(S, q^1, \ldots, q^n; x, t), \tag{2a}$$

$$q^j_t = B(S, q^1, \ldots, q^n; x, t), \tag{2b}$$

are integrable for all $j=1,\ldots,n$, subject to the constraint (1). [It is important to notice that A and B are in general nonlinear functions of the q^j; q^j itself in general appearing on the right-hand side of (2), as contrasted with potentials [1] which are determined via equations with no q^j on the right-hand side.] The correspondence between this definition and that of Wahlquist and Estabrook is discussed in Ref. 4.

A *pseudopotential of the first kind* is the restriction of the above to the case when $n=1$, and A and B are not explicitly functions of x and t. Hence a pseudopotential of the first kind is a function $q(x,t)$ such that

$$q_x = A(S,q), \tag{3a}$$

$$q_t = B(S,q) \tag{3b}$$

are integrable on (1).

In every case so far considered A and B have been found to be of the form

$$A = \sum_{k=1}^{I} A_k(S) X_k(q), \tag{4a}$$

$$B = \sum_{j=1}^{I'} B_j(S) X_j(q). \tag{4b}$$

In addition it is always found (and this must be so) that the $X_i(q)$ are determined by equations which involve terms of the form

$$\frac{\partial X_k}{\partial q} X_i - X_k \frac{\partial X_i}{\partial q} = [X_k, X_i]. \tag{5}$$

A pseudopotential of the first kind is called *Abelian* if all terms of the form (5) vanish and *non-Abelian* if at least one does not. Abelian pseudopotentials are equivalent to potentials.[11] A non-Abelian pseudopotential of the first kind will be called a *simple pseudopotential*.

In the example that follows, and in all other cases computed thus far, when a simple pseudopotential is found the X_r's form a Lie algebra under the bracket defined by (5). For pseudopotentials of the first kind the Lie algebra structure follows naturally and does not have to be forced by "assuming closure" as in Ref. 1.

III. THE HIROTA EQUATION

Consider

$$\varphi_t = -3\alpha\varphi\bar{\varphi}\varphi_x - \beta\varphi_{xxx} + i\gamma\varphi_{xx} + i\epsilon\varphi^2\bar{\varphi}, \tag{6}$$

where $\alpha, \beta, \gamma, \epsilon$ are real constants, φ is a scalar function function, and $\bar{\varphi}$ its complex conjugate. When $\alpha = \beta = 0$, this equation reduces to the nonlinear Schrodinger equation, while when $\gamma = \epsilon = 0$ the modified Korteweg–de Vries equation results.

The pseudopotential of the first kind associated with (6) is defined by the pair of equations

$$q_x = A(\varphi, \bar{\varphi}, z, \bar{z}, p, \bar{p}, q), \tag{7a}$$

$$q_t = B(\varphi, \bar{\varphi}, z, \bar{z}, p, \bar{p}, q), \tag{7b}$$

which are assumed integrable on (6). Here $z = \varphi_x$, $p = \varphi_{xx}$, $\bar{z} = \bar{\varphi}_x$, $\bar{p} = \bar{\varphi}_{xx}$. In this notation

$$\varphi_t = -3\alpha\varphi\bar{\varphi}z - \beta p_x + i\gamma p + i\epsilon\varphi^2\bar{\varphi}. \tag{8}$$

By imposing the integrability condition on (7) with (8) as a constraint is a standard computation, however, its outline may be of some use. For (7) to be integrable it is necessary and sufficient that $q_{xt} = q_{tx}$, or explicitly that

$$A_\varphi\varphi_t + A_{\bar{\varphi}}\bar{\varphi}_t + A_z z_t + A_{\bar{z}}\bar{z}_t + A_p p_t + A_q B + A_{\bar{p}}\bar{p}_t = B_\varphi z + B_{\bar{\varphi}}\bar{z} + B_z p + B_{\bar{z}}\bar{p} + B_p p_x + B_{\bar{p}}\bar{p}_x + B_q A, \tag{9}$$

where (7) has been used to replace q_t and q_x by A and B respectively. Now φ_t and $\bar{\varphi}_t$ can be replaced by the right-hand side of (8) and its complex conjugate. Likewise the derivative of (8) can be used to replace z_t. Similarly for $\bar{z}_t, p_t, \bar{p}_t$. Since p_t introduces terms involving p_{xxx} and there are no terms on the right-hand side of (9) that involve p_{xxx}, its coefficient A_p must vanish. Similar arguments show that $A = A(\varphi, \bar{\varphi}, q)$ and what remains of (9) is

$$A_\varphi\{-3\alpha\varphi\bar{\varphi}z - \beta p_x + i\gamma p + i\epsilon\varphi^2\bar{\varphi}\} + A_{\bar{\varphi}}\{-3\alpha\varphi\bar{\varphi}\bar{z} - \beta\bar{p}_x - i\gamma\bar{p} - i\epsilon\bar{\varphi}^2\bar{\varphi}\} + [A, B] = zB_\varphi + \bar{z}B_{\bar{\varphi}} + pB_z + \bar{p}B_{\bar{z}} + p_x B_p + \bar{p}_x B_{\bar{p}}, \tag{10}$$

where, following (5), $[A, B] = A_q B - AB_q$.

By noting that the dependence on p_x and $\bar{p}_x$ is explicit, it follows that

$$-\beta A\varphi = Bp, \tag{10a}$$

$$-\beta A\bar{\varphi} = B\bar{p} \tag{10b}$$

and

$$A_\varphi\{-3\alpha\varphi\bar{\varphi}z + i\gamma p + i\epsilon\varphi^2\bar{\varphi}\} + A_{\bar{\varphi}}\{-3\alpha\varphi\bar{\varphi}\bar{z} - i\gamma\bar{p} - i\epsilon\bar{\varphi}^2\varphi\} + [A, B] = zB_\varphi + \bar{z}B_{\bar{\varphi}} + pB_z + \bar{p}B_{\bar{z}}. \tag{11}$$

The (integrable) system (10) implies

$$B = -\beta pA_\varphi - \beta\bar{p}A_{\bar{\varphi}} + C(\varphi, \bar{\varphi}, z, \bar{z}, q), \tag{12}$$

where C is to be determined from (11).

If (12) is substituted into (11), the coefficients of $\bar{p}$ in the equation can be equated since all the p dependence is now explicit. The same is true of the coefficients of p. This balancing results in equations for C_z and $C_{\bar{z}}$, which can be integrated. The result is substituted into what remains of (10). Now coefficients of z^2, $\bar{z}$ and so on are separately equated and the process continues, the term carried from one step to the next rapidly gaining weight. The process is, however, algorithmic.

The net result is the determination of the explicit φ, $\bar{\varphi}$, $z, \bar{z}, p$ and $\bar{p}$ dependence of A and B plus a set of differential conditions on the q-dependent functions. The particular forms of A and B are of no interest for the moment. The specialization that results when (6) reduces to the nonlinear Schrödinger equation will be presented and used in the next section. It is sufficient to note that A and B are of the form (4). Of immediate interest is the relations among the X_r's.

At this point it is convenient to pass over a series of straightforward arguments and merely quote the conclusions. The deleted arguments show that the differential constraints of the X_r imply that a certain subset of them must be zero if the pseudopotential is to be non-Abelian. The details consist of following a fairly complicated logical tree, each step is, however, immediate. An example of an analogous type of argumentation is given below, where it is shown that if $X_1 = 0$ the resulting structure is Abelian.

With the above argument given the set of relations among the X_r is as follows

$$-\alpha X_1 + \beta[X_1, [X_1, X_2]] = 0, \tag{13a}$$

$$\alpha X_2 + \beta[X_2, [X_1, X_2]] = 0, \tag{13b}$$

$$[X_3, [X_1, X_2]] = 0, \tag{13c}$$

$$i\epsilon X_1 + i\gamma[X_1, [X_2, X_1]] - \beta[X_1, [X_2, [X_3, X_1]]] - \tfrac{1}{2}\beta[X_2, [X_1, [X_3, X_1]]] - \alpha[X_3, X_1] = 0, \tag{13d}$$

$$-i\epsilon X_2 - i\gamma[X_2, [X_2, X_1]] - \beta[X_2, [X_2, [X_3, X_1]]] - \tfrac{1}{2}\beta[X_1, [X_2, [X_3, X_2]]] - \alpha[X_3, X_2] = 0, \tag{13e}$$

$$-\tfrac{1}{2}\beta[X_1, [X_1, [X_3, X_1]]] = 0, \tag{13f}$$

$$-\tfrac{1}{2}\beta[X_2, [X_2, [X_3, X_2]]] = 0, \tag{13g}$$

$$-i\gamma[X_1, [X_1, X_3]] - \beta[X_3, [X_1, [X_3, X_1]]] - \beta[X_1, [X_3, [X_3, X_4]]] = 0, \tag{13h}$$

$$-i\gamma[X_2, [X_2, X_3]] - \beta[X_2, [X_3, [X_3, X_2]]] - \tfrac{1}{2}\beta[X_3, [X_2, [X_3, X_2]]] = 0, \tag{13i}$$

$$-i\gamma[X_1, [X_2, X_3]] + i\gamma[X_2, [X_3, X_1]] + i\gamma[X_3, [X_2, X_1]] - \beta[X_1, [X_3, [X_3, X_2]]] - \beta[X_2, [X_3, [X_3, X_1]]] - \beta[X_3, [X_2, [X_3, X_1]]] + [X_5, X_4] = 0, \tag{13j}$$

$$-i\gamma[X_3, [X_1, X_3]] - \beta[X_3, [X_3, [X_3, X_1]]] + [X_1, X_4] = 0, \tag{13k}$$

$$i\gamma[X_3, [X_2, X_3]] - \beta[X_3, [X_3, [X_3, X_2]]] + [X_2, X_4] = 0, \tag{13l}$$

$$[X_3, X_5] = 0. \tag{13m}$$

The system (13) in a formidable array of conditions. There are but five functions related by thirteen conditions. However, the prospects of obtaining a solution are not as bleak as it might first appear. This is due to two facts which are direct consequences of the restriction to pseudopotentials of the first kind.

The first essential observation is that if

$$q_x = \sum_{i=1}^{I} A_i(S)X_i(q) \tag{14}$$

and it is known that, say, X_1 is nonzero, then in (14) it can be assumed, without loss of generality, that $X_1=1$. This can be done more elaborately by dividing through by X_1 and introducing new variables. The relevant transformation properties of pseudopotentials are discussed in Ref. 4. If different forms of the explicit pseudopotential that results at the end of the computation are needed, $q=f(\tilde{q})$ can be introduced.

The second essential observation is that if $[X_i, X_j] = 0$, then $X_i = aX_j$, where a is a constant. Again this is due to the fact that the system (13) is a set of ordinary differential equations since pseudopotentials of the first kind are being computed.

So, for example, (13c) shows that

$$[X_1, X_2] = aX_3. \tag{15}$$

Simple inspection of (13) shows that if $X_3=0$, the remaining structure is Abelian. Hence, it is possible to set $X_3=1$. With this assignment (13a), (13b), and (15) can easily be solved consistently, at each stage a linear first order ordinary differential equation is solved. The result is

$$X_1 = b\exp(\alpha/a\beta)q, \tag{16a}$$

$$X_2 = c\exp(-\alpha/a\beta)q, \tag{16b}$$

$$a^2 = 2\alpha/\beta. \tag{16c}$$

If these results are used in (13d) or (13e), it follows that

$$(\beta\epsilon - \gamma\alpha)X_1 = 0. \tag{17}$$

If $X_1=0$, (13a) implies $X_2=0$ and (13k) shows $X_4=a'X_5$ while (13m) implies $X_3=a''X_5$, and an Abelian structure results. Hence a necessary condition for the pseudopotential to be non-Abelian is

$$(\beta\epsilon - \gamma\alpha) = 0. \tag{18}$$

This is precisely the condition derived by Hirota in order that (6) have n-soliton solutions.

The calculations can easily be carried through in order to obtain X_4 and X_5 explicitly. No additional constraints on the form of (6) are obtained. It is interesting to note that the set $\{X_1, X_2, X_3, X_4, X_5\}$ has the property that

$$[X_i, X_j] = \sum_k a_{ijk}X_k \tag{19}$$

for all i,j. That is, it is a Lie algebra under the bracket defined by (5). All non-Abelian pseudopotentials of the first kind that have been computed define a Lie algebra in this way. Notice that this structure follows in a natural way from the construction. No "closure" assumption, such as used in Ref. 1, is needed to force the structure.

IV. BÄCKLUND TRANSFORMATIONS

Simple pseudopotentials are useful in their own right. In particular they can be used to find a "new" solution of of an equation in terms of another (or "old") solution. The computation is again straightforward. A useful example is provided by the nonlinear Schrödinger equation.

Consider the specialization of (6) to

$$i\varphi_t = -\varphi_{xx} + k\bar{\varphi}\varphi^2. \tag{20}$$

The simple pseudopotential associated with this equation can easily be found by the method of the previous section. The result is

$$q_x = \tfrac{1}{2}ik\varphi q^2 + i\bar{\varphi} - i\lambda q, \tag{21a}$$

$$q_t = -\tfrac{1}{2}kzq^2 + \bar{z} + ik\varphi\bar{\varphi}q - \tfrac{1}{2}i\lambda k\varphi q^2 - i\lambda\bar{\varphi} + i\lambda^2 q. \tag{21b}$$

Two remarks are in order. First (21) is called *the* simple psedopotential since it is understood, as indicated earlier, that all pseudopotentials that can be reached via the coordinate transformation $q=f(\tilde{q})$ are equivalent. Second, if a computation analogous to that done in the previous section is attempted based on (20), nontrivial results are obtained only if k is real.

Consider now a function

$$\psi = \psi(\varphi, q, \bar{q}). \tag{22}$$

If it is required that ψ is a solution of (20) provided φ is a solution of (20) and q and $\bar{q}$ satisfy (21) and its complex conjugate, a straightforward calculation shows

$$\psi = \varphi + 2(\lambda - \bar{\lambda})\bar{q}/(2 - kq\bar{q}), \tag{23}$$

as can easily be verified.

If now (23) and its complex conjugate are used to express q in terms of $\varphi, \psi, \bar{\varphi}$, and $\bar{\psi}$ and the result substituted into (21), the Bäcklund transformation[3] for (20) results.

The same methods yields the Bäcklund transformation for the KdV and sine–Gordon equations.[4] Again a pseudopotential of the first kind is all that is needed to obtain Bäcklund transformations.

The above method was suggested by the work of Wahlquist and Estabrook. It is, however, more direct. While this work was in progress Wahlquist and Estabrook showed, in a preprint,[5] how to obtain (essentially) (23) using a very interesting argument based on Lie derivatives of the ideal of forms which represent (20).

V. CONCLUSION

With which equations can simple pseudopotentials be associated? The results of Sec. III show that the existence of simple pseudopotentials is extremely sensitive to the detailed structure of the equation which is the starting point of the computation. Indeed the result suggests that there is a connection between the existence of simple pseudopotentials and the existence of solitons. There is additional evidence which points to this conclusion.

In particular, it can be shown[6] that, for a simple pseudopotential to be associated with

$$\varphi_t + f(\varphi)\varphi_x + \varphi_{xxx} = 0, \tag{24}$$

it is necessary that

$$f_{\varphi\varphi\varphi} = 0. \tag{25}$$

Wahlquist has shown,[7] that this condition is necessary if any pseudopotential is to exist. In addition, for a

simple pseudopotential to be associated with

$$\varphi_{xt} = g(\varphi), \tag{26}$$

it is necessary that[4]

$$g_{\varphi\varphi} = ag \text{ a constant.} \tag{27}$$

It is known from numerical studies that[8]

$$\varphi_t + \varphi^3\varphi_x + \varphi_{xxx} = 0$$

does not possess soliton solutions. It is also known that equations of the form (24) only possess an infinite number of (polynomial) conservation laws if (25) is satisfied.[9] The condition (27) was first derived by Kruskal as a condition for (26) to have an infinite number of conservation laws[10] and was rederived in Ref. 11 as a condition for (27) to possess a Bäcklund transformation.

Taking the evidence in sum, it is not unreasonable to attempt to show that, for an evolution equation to have soliton solutions, it is necessary that a simple pseudopotential be associated with the equation. Nonlinear hyperbolic equations, such as (26), probably require a different criterion, however. It is emphasized that the question of sufficient conditions for the existence of solitons has not been addressed.

It is clear that the Lax criterion,[12] which has been and must continue to be at the center of any studies which actually find solutions of equations of the form (1) by the inverse method,[13,14] is only useful if pairs of operators which satisfy the criterion have somehow been found. There is no known method for constructing them and no results which state when they can in principle be found. In addition it would appear[4] that satisfying the criterion *alone* is only necessary for the existence of solitons, although this extremely important point has not been treated in the literature. A careful study of pseudopotentials, both simple and nonsimple, could be undertaken with the realistic hope of obtaining *computable* criteria which must be satisfied if soliton solutions to a particular equation exist.

[1]H. D. Wahlquist and F. B. Estabrook, J. Math. Phys. **16**, 1 (1975).
[2]R. Hirota, J. Math. Phys. **14**, 805 (1973).
[3]G. Lamb, J. Math. Phys. **15**, 2157 (1974).
[4]J. P. Corones and F. J. Testa, to appear in the Proceedings of the NSF Workshop on Contact Transformations, Vanderbilt University, September 1974.
[5]F. B. Estabrook and H. D. Wahlquist, "Prolongation Structures of Nonlinear Evolution Equations. II. The Nonlinear Schroedinger Equation."
[6]J. P. Corones, unpublished.
[7]H. D. Wahlquist, private communication (1975).
[8]N. J. Zabusky, Comput. Phys. Commun. **5**, 1 (1973).
[9]R. M. Miura, in *Nonlinear Waves*, edited by Leibovich and Seebass (Cornell U. P., Itaca, N.Y., 1974).
[10]R. M. Miura, private communication.
[11]D. W. McLaughlin and A. C. Scott, J. Math. Phys. **14**, 1817 (1973).
[12]P. D. Lax, Comm. Pure Appl. Math. **21**, 467 (1968).
[13]C. S. Gardner, J. M. Greene, M. D. Kruskal, and R. M. Miura, Phys. Rev. Lett. **19**, 1095 (1967).
[14]M. J. Ablowitz, D. J. Kaup, A. C. Newell, and H. Segur, Studies Appl. Math. **53**, 249 (1974).

Asymptotic approximations to angular-spectrum representations*

George C. Sherman, Jakob J. Stamnes,† and Éamon Lalor‡

The Institute of Optics, The University of Rochester, Rochester, New York 14627
(Received 29 April 1975)

Under rather general conditions, a time-harmonic wave field $u(x, y, z)$ can be represented in a half-space $z > 0$ by a double integral known as the angular spectrum of plane waves. The representation divides naturally into the sum of two double integrals, one of which (u_H) is a superposition of homogeneous plane waves and the other (u_I) is a superposition of inhomogeneous plane waves. We obtain asymptotic approximations to $u(x, y, z)$, u_H, and u_I valid when the point of observation of the field recedes towards infinity in a fixed direction through a fixed point. The results apply when the spectral amplitude of the plane waves belongs to a specific class which arises frequently in applications. Our approach is based on the method of stationary phase, which we extend in order to permit the presence of inhomogeneous waves in the integrand. Although the analysis of u requires that we distinguish the directions that are perpendicular to the z axis from the directions pointing into the half-space $z > 0$, the results for the former case are the same as would be obtained by taking the appropriate limit in the results of the latter case. We obtain the general form of the asymptotic sequence appropriate for expanding u and present explicit expressions for the first two terms. Our derivation justifies the results of previous heuristic treatments. The analysis of u_H and u_I requires separate treatments for directions that are (i) perpendicular to the z axis, (ii) parallel to the z axis, and (iii) neither perpendicular nor parallel to the z axis. In contrast to the behavior of u, the asymptotic behavior of u_H (and of u_I) differs in the different cases. In each case, we obtain the general form of the appropriate asymptotic sequence and present the first term explicitly.

1. INTRODUCTION

A. Statement of the problem

In the theories of acoustics, electrodynamics, and physical optics, the complex amplitude $u(x, y, z)$ of a monochromatic scalar wave field $u(x, y, z, t) = u(x, y, z) \times \exp(-i\omega t)$ is frequently expressed in the half-space $z > 0$ by the angular spectrum of plane-waves representation

$$u(x, y, z) = u_H(x, y, z) + u_I(x, y, z) \quad \text{for } z > 0, \tag{1.1}$$

where

$$u_J(x, y, z) = \int\int_{D_J} U(p, q) \exp[ik(px + qy + mz)]\, dp\, dq \qquad (J = H, I), \tag{1.2}$$

D_H is the region $0 \leq p^2 + q^2 \leq 1$,

D_I is the region $p^2 + q^2 \geq 1$,

$$m = +(1 - p^2 - q^2)^{1/2} \quad \text{in } D_H, \tag{1.3a}$$

$$m = +i(p^2 + q^2 - 1)^{1/2} \quad \text{in } D_I. \tag{1.3b}$$

k is a real positive constant, and $U(p, q)$ is a spectral amplitude function, independent of x, y, z, that characterizes the field $u(x, y, z)$. With sufficiently well-behaved spectral amplitude $U(p, q)$, $u(x, y, z)$ satisfies the homogeneous Helmholtz equation

$$\nabla^2 u + k^2 u = 0, \tag{1.4}$$

and the Sommerfeld radiation condition in the half-space $z > 0$.[1,2] In (1.2), $u_H(x, y, z)$ is expressed as a superposition of homogeneous plane waves propagating in different directions; $u_I(x, y, z)$ is expressed as a superposition of inhomogeneous plane waves, propagating in directions parallel to the plane $z = 0$ and decaying at different rates in the z direction. Because the properties of the inhomogeneous plane waves differ greatly from those of the homogeneous plane waves, $u_H(x, y, z)$ and $u_I(x, y, z)$ are usually treated separately in the mathematical analysis and its physical interpretation.

Since the integral in (1.2) can rarely be evaluated analytically, approximations of the integral are required. The approximations of greatest importance for most applications are those valid at large distances

$$R = [(x - x_0)^2 + (y - y_0)^2 + (z - z_0)^2]^{1/2}, \tag{1.5}$$

from a point (x_0, y_0, z_0). In this paper, we study the asymptotic behavior of $u(x, y, z)$, $u_H(x, y, z)$, and $u_I(x, y, z)$ for large kR with fixed k and fixed direction cosines ξ_1, ξ_2, ξ_3 given by

$$\xi_1 = (x - x_0)/R, \tag{1.6}$$

$$\xi_2 = (y - y_0)/R, \tag{1.7}$$

$$\xi_3 = (z - z_0)/R. \tag{1.8}$$

We obtain in Sec. 2, an asymptotic approximation of $u(x, y, z)$ valid when the point of observation (x, y, z) recedes towards infinity in a fixed direction with positive z component ($\xi_3 > 0$) through a point (x_0, y_0, z_0). In Sec. 3, we obtain an asymptotic approximation of $u(x, y, z)$ valid when the point of observation recedes towards infinity in a fixed direction perpendicular to the z axis ($\xi_3 = 0$) through a point (x_0, y_0, z_0) in the half-space $z > 0$. The dominant terms in the asymptotic expansions of $u_H(x, y, z)$ and $u_I(x, y, z)$ are obtained in Sec. 4. Our main results are summarized in Sec. 5.

In order to obtain the asymptotic approximations, some restrictions must be placed on the spectral amplitude $U(p, q)$. We consider spectral amplitudes that belong to a set T_N defined for positive, even integer N as the set of all functions $U(p, q)$ that are independent of x, y, and z, and that satisfy the conditions:

(i) $U(p, q)$ can be written in the form [with m given by (1.3)]

$$U(p, q) = V(p, q, m)/m, \tag{1.9}$$

where $V(p, q, m)$ is bounded for all p, q;

(ii) $V(p, q, s)$ is a real or complex, continuous function of three independent variables p, q, s defined (a) for all real p and q, (b) for real s such that $0 \leq s \leq 1$, and (c) for purely imaginary $s = i\sigma$ such that $\sigma > 0$;

(iii) $V(p, q, s)$ has continuous, bounded partial derivatives up to order N with respect to p, q, s for all p, q, s within its domain of definition.

Wave fields with spectral amplitudes $U(p, q)$ that belong to T_N arise frequently in the theories of radiation and diffraction of waves. For example, consider the wave field radiated by a time-harmonic source of finite strength and size located in the region $z \leq 0$. The complex amplitude $u(x, y, z)$ of the field satisfies the inhomogeneous Helmholtz equation

$$\nabla^2 u + k^2 u = -4\pi\rho(x, y, z), \tag{1.10}$$

where $\rho(x, y, z)$ is bounded and vanishes outside the region occupied by the source. Then $u(x, y, z)$ is given by (1.1) with[3]

$$U(p, q) = (ik/2\pi m)\tilde{\rho}(kp, kq, km), \tag{1.11}$$

where $\tilde{\rho}(k_x, k_y, k_z)$ is the three-dimensional Fourier transform of $\rho(x, y, z)$ defined by

$$\tilde{\rho}(k_x, k_y, k_z) = \int_{-\infty}^{\infty}\int\int \rho(x, y, z) \exp[-i(k_x x + k_y y + k_z z)]\, dx\, dy\, dz. \tag{1.12}$$

Since $\rho(x, y, z)$ vanishes outside some bounded region, it follows from (1.12) that $\tilde{\rho}(k_x, k_y, k_z)$ is an entire function of complex k_x, k_y, and k_z. Furthermore, since $\rho(x, y, z)$ vanishes for $z > 0$, it follows from (1.12) that $\tilde{\rho}(kp, kq, km)$ and all of its partial derivatives are bounded for real p, q. Consequently, $U(p, q)$ belongs to T_N for arbitrarily large N.

Another example of a wave field with spectral amplitude in T_N, where N is arbitrarily large, is a field $u(x, y, z)$ expressible by (1.1) in $z \geq 0$ with bounded boundary value $u(x, y, 0)$ that vanishes outside a bounded region of the plane $z = 0$. It may be readily verified that in this case $V(p, q, m)$ in (1.9) is of the form

$$V(p, q, m) = mW(p, q, m), \tag{1.13}$$

where $W(p, q, s)$ has the same properties as $V(p, q, s)$ specified in conditions (ii)—(iii) above. In fact,

$$W(p, q, m) = (k/2\pi)^2 \tilde{u}(kp, kq), \tag{1.14}$$

where $\tilde{u}(k_x, k_y)$ is the two-dimensional Fourier transform of $u(x, y, 0)$ defined in analogy to the definition of the three-dimensional Fourier transform in (1.12).

Although in both examples just cited, $V(p, q, s)$ is an entire function of complex p, q, and s, we do not restrict our analysis here to such functions. The benefits of dealing with more general $U(p, q)$ is pointed out in Sec. 5. Also in Sec. 5, we comment on examples of spectral amplitudes that arise in physical problems but that are not in T_N, and we indicate how they can be handled.

Condition (i) on the functions in T_N guarantees the convergence of the integrals in (1.2) for $U(p, q) \in T_N$ and $z > 0$. Conditions (ii) and (iii) are needed to obtain the asymptotic approximations to follow.

B. Discussion of the problem

In order to gain insight into the asymptotic behavior of $u(x, y, z)$, let us discuss its integral representation from a heuristic point of view. (For examples of the application of this point of view to obtain the asymptotic behavior of $u(x, y, z)$ in a variety of physical problems see Refs. 4 and 5.)

When $\xi_3 > 0$, the inhomogeneous plane waves in $u_I(x, y, z)$ decay exponentially with increasing kz. Hence, we ignore $u_I(x, y, z)$ compared to $u_H(x, y, z)$ and take the asymptotic behavior of $u(x, y, z)$ to be the same as that of $u_H(x, y, z)$. Then, since the argument of the exponential in the integrand is purely imaginary in D_H, we obtain the asymptotic expansion of $u_H(x, y, z)$ by applying the method of stationary phase.

The method of stationary phase for double integrals is a technique for obtaining the asymptotic behavior for large kR of integrals of the form

$$I(kR) = \int\int_D g(p, q) \exp[ikRf(p, q)]\, dp\, dq, \tag{1.15}$$

where $g(p, q)$, and $f(p, q)$ are independent of kR and are sufficiently smooth in D, and where $kRf(p, q)$ is real in D. The heuristic basis for the method is very simple. For large enough kR, the rapid oscillation of the exponential resulting from small variations in p, q leads to a cancellation effect so that most of the domain D contributes only a negligible amount to $I(kR)$. The important contributions to $I(kR)$ arise from the neighborhoods of certain critical points. In particular, a point p_s, q_s (called an interior stationary point) within the interior of D where the phase $f(p, q)$ is stationary (i.e., where both $\partial f/\partial p$ and $\partial f/\partial q$ vanish) is such a critical point because the exponential does not oscillate there even for large kR. Other types of critical points can occur on the boundary and at locations of singularities of the integrand.

It is difficult, however, to apply these heuristic ideas of stationary phase to develop a rigorous theory for the asymptotic behavior of integrals in the form of (1.15). Consequently, other approaches are used. For example, Van der Corput[6] and Erdélyi[7] apply integration by parts to treat the corresponding single integrals. Rigorous treatments of the double integral $I(kR)$ under various conditions have been given by Focke,[8] Braun,[9] Jones and Kline,[10] and Chako.[11] Although the approaches used in the above references do not resemble the heuristic argument of stationary phase, the analyses therein supply a rigorous justification of the results of that argument provided certain restrictions are placed on $f(p, q)$ and $g(p, q)$. Consequently the term "method of stationary phase" is often applied to describe the rigorous treatments. In this paper, we adopt that terminology and denote by the MSP for double integrals the combined methods of Focke,[8] Braun,[9] Jones and Kline,[10] and Chako.[11]

The MSP requires that, throughout the interior and boundary of D, $f(p, q)$ and $g(p, q)$ have at least a finite number of continuous partial derivatives except possibly

at a finite number of isolated points where $g(p, q)$ may have integrable singularities of a certain type. Unfortunately, integrals of interest in many applications do not satisfy these conditions. In particular, the integrals treated in this paper do not satisfy the conditions on the circle $p^2+q^2=1$. Neither $f(p, q)$ nor $g(p, q)$ is differentiable there. Moreover, if $V(p, q, m)$ is not zero for $p^2+q^2=1$, the integrand is infinite on a continuum of points (the unit circle) rather than at isolated points. Hence, the circle $p^2+q^2=1$ must be dealt with in some way if we are to apply the MSP to obtain asymptotic expansions of $u(x, y, z), u_H(x, y, z)$, and $u_I(x, y, z)$.

Another important requirement of the MSP is that $kRf(p, q)$ be real in D. Consequently, the MSP cannot be applied to $u(x, y, z)$ or $u_I(x, y, z)$ when $\xi_3 > 0$ since m is then imaginary in D_I. Of course, if $u_I(x, y, z)$ can be neglected compared to $u_H(x, y, z)$ in $u(x, y, z)$ as suggested in the heuristic discussion earlier, then at least the first term in the asymptotic behavior of $u(x, y, z)$ can be obtained from the asymptotic behavior of $u_H(x, y, z)$. It is easy to see, however, that the heuristic argument breaks down in certain cases and that $u_I(x, y, z)$ can be just as important as $u_H(x, y, z)$ for large kR. To that end, consider the spherical wave $u^s(x, y, z)$ radiating from the origin, i.e.,

$$u^s(x, y, z) = \exp[ik(x^2+y^2+z^2)^{1/2}]/(x^2+y^2+z^2)^{1/2}. \tag{1.16}$$

For this wave field, the spectral amplitude is (Ref. 12, Sec. 2.11)

$$U^s(p, q) = ik/2\pi m. \tag{1.17}$$

When the point of observation lies on the z axis, the integrals can be evaluated easily to obtain

$$U_H^s(x, y, z) = \exp(ikz)/z - 1/z, \tag{1.18}$$

$$U_I^s(x, y, z) = 1/z. \tag{1.19}$$

Hence, in this case, $u_H(x, y, z)$ and $u_I(x, y, z)$ are of the same order in $1/z$ so that it is incorrect to neglect $u_I(x, y, z)$ compared to $u_H(x, y, z)$. As we will see later in the paper, this rather surprising behavior is a consequence of the singularity on the circle $p^2+q^2=1$ due to the factor $1/m$ in (1.17). For $\xi_3=1$, $u_I(x, y, z)$ is of the same order in $1/R$ as is $u(x, y, z)$ for all $U(p, q) \in T_N$ with nonvanishing $V(p, q, 0)$.

Since $u_I(x, y, z)$ cannot be neglected compared to $u_H(x, y, z)$ in general, we are forced to deal with integrals of the form of $I(kR)$ in (1.15) with complex $f(p, q)$ in order to obtain the asymptotic behavior of $u(x, y, z)$. For such an integral, the MSP is not applicable. A method that is available for obtaining asymptotic expansions of single integrals analogous to $I(kR)$ in (1.15) with complex $f(p, q)$ is the method of steepest descents (MSD). An important requirement of the method is that the integrand be an analytic function of the variable of integration for complex values of that variable. This method has been applied, for example, in Chaps. 4–7 of Ref. 12, Chaps. 4–5 of Ref. 13, and in Chaps. 6–7 of Ref. 14, to treat some special cases in which the representation of $u(x, y, z)$ reduces to a single integral.

In many cases of interest, $U(p, q)$ is an analytic function of complex p, q, but even in these cases, we cannot apply the MSD directly to our integrals. The MSD has not yet been developed for multiple integrals. A heuristic approach to the method for multiple integrals has been given by Baños[15] to obtain the first-order term, but he points out that his approach cannot be extended to obtain higher order terms or to make estimates of the remainder term. The only MSD approach available at present to treat a double integral of the form in (1.15), is to treat it as an iterated integral and to apply the MSD to each single integral successively. This approach is very cumbersome, and no general results are available.

Morse and Feshbach[16] apply a change of variables of integration to place the integral $u(x, y, z)$ in the form of $I(kR)$ in (1.15) with $f(p, q)$ a function of p only. They then apply the MSD to obtain the first term in the asymptotic expansion of the p integral. Since the result is independent of q, the q integral can be done immediately. Unfortunately, the domain D of integration in the integral obtained by the change of integration variables is a complicated two-dimensional surface in the four-dimensional space of two complex variables. Morse and Feshbach do not determine this domain D nor do they discuss how it can be transformed into the domain of integration they use when they apply the MSD. Further analysis is required to resolve this difficulty.

C. Present approach to the problem

In Sec. 1, Part B., several possible methods for obtaining the desired asymptotic approximations are discussed, and the difficulties associated with each are pointed out. A logical approach to the problem is to select one of the methods and to attempt to resolve the difficulties associated with it. It is not *a priori* obvious which of these methods can be rectified most readily. It is clear, however, that the methods based on the MSD have the following additional disadvantages over those based on the MSP:

(a) they cannot be used for all $U(p, q) \in T_N$ since they require that $U(p, q)$ be an analytic function of complex p, q;

(b) they cannot be used to study $u_H(x, y, z)$ and $u_I(x, y, z)$ separately;

(c) they require modification of the contour of integration so that p and q become complex and the physical interpretation of the interference of homogeneous plane waves associated with the MSP is lost.

For these reasons, the approach taken in this paper is to apply the MSP, extending it where necessary to avoid the difficulties mentioned in Sec. 1, Part B.

In Sec. 2, where we treat the case $\xi_3 > 0$, we employ the neutralizer functions introduced by Van der Corput[6] to isolate a neighborhood of the circle $p^2+q^2=1$ and apply an argument based on integration by parts to show that that neighborhood does not contribute to those terms in the asymptotic expansion of $u(x, y, z)$ that are of order lower than $(kR)^{-N}$. The exponential decay of the inhomogeneous waves is utilized to show that the integral over D_I, with the neighborhood of the unit circle excluded by

the neutralizer, does not contribute to the asymptotic expansion of $u(x,y,z)$ in inverse powers of kR to any order. It follows then from Ref. 11 that the total contribution to the asymptotic expansion of $u(x,y,z)$ up to order $(kR)^{-N}$ is due to a neighborhood (again isolated by a neutralizer) of the interior stationary point $p=\xi_1$, $q=\xi_2$. This result provides a rigorous justification of the results yielded by the heuristic approach described in Sec. 1, Part B. The form of the asymptotic expansion complete with the order of the remainder is provided by Braun.[9] The first- and second-order terms are calculated explicitly using the expressions of Jones and Kline.[10]

The case $\xi_3=0$ treated in Sec. 3 is simplified by the fact that the integral representation of $u(x,y,z)$ is of the form in (1.15) with real $f(p,q)$, but it is complicated by the fact that the point of stationary phase lies on the circle $p^2+q^2=1$. We again employ a neutralizer to isolate the unit circle and show that the portion of the integral that does not include a neighborhood of $p^2+q^2=1$ does not contribute to those terms in the asymptotic expansion of $u(x,y,z)$ which are of order lower than $(kR)^{-N}$. In order to treat the integral over the neighborhood of $p^2+q^2=1$, a change of variable of integration is made to eliminate the singularities on $p^2+q^2=1$. The integral must be split into two parts for this purpose, since the new integration variables required in D_I differ from those required in D_H. Several new critical points are introduced by the change of variables, but we are able to treat them by judicious use of neutralizers and the MSP. We apply Ref. 9 to obtain the form of the asymptotic expansions and the error term and apply Ref. 10 to obtain explicit expressions for the coefficients. It is finally concluded that the asymptotic expansion of $u(x,y,z)$ for $\xi_3=0$, can be obtained by setting $\xi_3=0$ in the results of Sec. 2.

The integrals $u_H(x,y,z)$ and $u_I(x,y,z)$ are treated in Sec. 4. Here, the cases (a) $0<\xi_3<1$, (b) $\xi_3=0$, (c) $\xi_3=1$ must be treated separately. Whereas the asymptotic behavior of $u(x,y,z)$ is found to be the same in the three cases cited above, the asymptotic behavior of $u_J(x,y,z)$ $(J=H,I)$ is found to differ in the different cases. $u_I(x,y,z)$ is found to be of higher order in $(kR)^{-1}$ only in case (a).

A summary of the results written especially for the reader uninterested in the details of derivation, and a brief discussion on the application of our results in physical problems are presented in Sec. 5.

2. APPROXIMATION VALID OVER A HEMISPHERE

In this section, we obtain an asymptotic approximation of $u(x,y,z)$ valid as the point of observation recedes towards infinity in a fixed direction with positive z-component through a point (x_0,y_0,z_0). To do that, we first extend the MSP in Sec. 2, Part A., to show that when $U(p,q)\in T_N$, only an arbitrarily small neighborhood of the interior stationary point $p_s=\xi_1$, $q_s=\xi_2$ contributes to the asymptotic behavior up to terms of the order of $(kR)^{-N}$. We then apply, in Sec. 2, Part B., standard results of the MSP to obtain the asymptotic contribution due to the stationary point.

A. Extension of the MSP

Following the approach of Focke and Chako, we isolate the interior stationary point using a neutralizer function $\nu(p,q)$. To construct $\nu(p,q)$, we note that since the point of observation recedes towards infinity in a fixed direction with positive z-component, we have constant ξ_1, ξ_2, ξ_3 with $\xi_3>0$. As a result, the point $p_s=\xi_1$, $q_s=\xi_2$ is fixed within the interior of D_H. Let Ω_1 and Ω_2 be neighborhoods (arbitrarily small) of p_s, q_s that lie completely within the interior of D_H, and let Ω_1 be a subset of Ω_2. Then we take $\nu(p,q)$ to be a real, continuous function of p,q with continuous partial derivatives of all orders for all real p,q. Moreover, we require

$$0\le\nu(p,q)\le 1, \tag{2.1}$$

$$\nu(p,q)=1 \qquad \text{for } p,q \text{ in } \Omega_1, \tag{2.2}$$

$$\nu(p,q)=0 \qquad \text{for } p,q \text{ not in } \Omega_2. \tag{2.3}$$

Such functions exist for arbitrary Ω_1 and Ω_2.[17] Additional details of the explicit form of $\nu(p,q)$ are unimportant here.

We are now in a position to state the primary result of this section.

Theorem: Let $U(p,q)\in T_N$ for some positive even integer N. Then for $z>0$, $u(x,y,z)$ given in (1.1) with k a positive constant satisfies

$$u(x,y,z)=u_0(x,y,z)+R(x,y,z), \tag{2.4}$$

where

$$u_0(x,y,z)=\int\int_{D_H}\nu(p,q)\exp[ik(px+qy+mz)]\,dp\,dq, \tag{2.5}$$

and

$$R(x,y,z)=O[(kR)^{-N}] \text{ as } kR\to\infty \tag{2.6}$$

uniformly with respect to ξ_1 and ξ_2 for all real ξ_1, ξ_2 such that $\delta<(1-\xi_1^2-\xi_2^2)^{1/2}\le 1$ for any positive constant $\delta<1$. By the statement that $R(x,y,z)$ satisfies (2.6) uniformly with respect to ξ_1, ξ_2, we mean that $R(x,y,z)$ is bounded above by $(kR)^{-N}$ multiplied by a constant that is independent of ξ_1, ξ_2 for all real ξ_1, ξ_2 such that $\delta<(1-\xi_1^2-\xi_2^2)^{1/2}\le 1$ for any positive constant $\delta<1$. It follows directly from the theorem that the asymptotic behavior of $u(x,y,z)$ of order lower than $(kR)^{-N}$ is determined completely by an arbitrarily small neighborhood of the point p_s, q_s. In particular, if $U(p,q)\in T_N$ where N can be arbitrarily large, then the complete asymptotic expansion of $u(x,y,z)$ is equal to the aysmptotic expansion of $u_0(x,y,x)$.

Our proof of the theorem is a straightforward modification of the proof of Theorem 1 in Ref. 11 or Theorem III in Ref. 8. We construct three new neutralizer functions $\nu_1(p,q)$, $\nu_2(p,q)$, and $\nu_3(p,q)$ all of which are real continuous functions of p,q with continuous partial derivatives of all orders for all p,q and which satisfy

$$0\le\nu_j(p,q)\le 1 \quad \text{for } j=1,2,3, \tag{2.7}$$

$$\nu_1(p,q)+\nu_2(p,q)+\nu_3(p,q)=1. \tag{2.8}$$

To specify the neutralizers further, we choose positive constants C_1,C_2,C_3,C_4 such that

$$(\xi_1^2+\xi_2^2)^{1/2}<C_1<C_2<1<C_3<C_4, \tag{2.9}$$

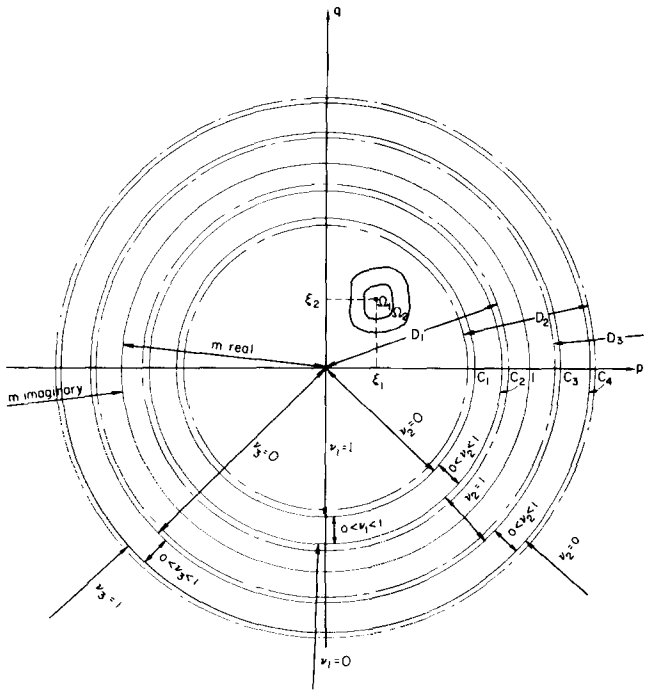

FIG. 1. Illustration of the notation used in Eqs. (2.7)—(2.17). The radii of the solid circles are marked off on the p axis. The distance between a dashed circle and the nearest solid circle is ϵ.

with C_1 large enough so that the neighborhood Ω_2 of the point p_s, q_s lies completely within the region $(p^2+q^2)^{1/2} < C_1$. Such constants exist since $\xi_1^2+\xi_2^2=1-\xi_3^2$ and $\xi_3 > \delta > 0$. Now we require that

$$\nu_1(p,q)=1 \quad \text{for} \quad (p^2+q^2)^{1/2} \leq C_1, \tag{2.10}$$

$$\nu_1(p,q)=0 \quad \text{for} \quad (p^2+q^2)^{1/2} \geq C_2, \tag{2.11}$$

$$\nu_2(p,q)=1 \quad \text{for} \quad C_2 \leq (p^2+q^2)^{1/2} \leq C_3, \tag{2.12}$$

$$\nu_2(p,q)=0 \quad \text{for} \quad (p^2+q^2)^{1/2} \leq C_1 \quad \text{and} \quad (p^2+q^2)^{1/2} \geq C_4, \tag{2.13}$$

$$\nu_3(p,q)=1 \quad \text{for} \quad (p^2+q^2)^{1/2} \geq C_4, \tag{2.14}$$

$$\nu_3(p,q)=0 \quad \text{for} \quad (p^2+q^2)^{1/2} \leq C_3. \tag{2.15}$$

Next, we define three regions D_1, D_2, D_3 in the p, q plane by:

D_1 is the region $(p^2+q^2)^{1/2} \leq C_2+\epsilon$,

D_2 is the region $C_1-\epsilon \leq (p^2+q^2)^{1/2} \leq C_4+\epsilon$,

D_3 is the region $C_3-\epsilon \leq (p^2+q^2)^{1/2}$,

where ϵ is a positive constant such that $\epsilon < 1-C_2$, $\epsilon < C_3-1$, and $\epsilon < C_1-(\xi_1^2+\xi_2^2)^{1/2}$. The various regions of interest in the p, q plane are shown in Fig. 1. The features important for the proof to follow are: (a) $\nu_j(p,q)$ vanishes for p, q not in D_j, and for p, q in D_j in some neighborhood of its boundary (with $j=1, 2, 3$); (b) the point p_s, q_s lies within the interior of D_1 and is exterior to D_2 and D_3; (c) m is real in D_1; and (d) m is imaginary in D_3. We are now ready to separate the integral $u(x,y,z)$ into three parts,

$$u(x,y,z)=u_1(x,y,z)+u_2(x,y,z)+u_3(x,y,z), \tag{2.16}$$

where

$$u_j(x,y,z)=\int\int_{D_j} \nu_j(p,q)\cdot U(p,q) \times \exp[ik(px+qy+mz)]\,dp\,dq, \tag{2.17}$$

and deal with each part separately.

The integral representation of $u_1(x,y,z)$ is in a form appropriate for the MSP. Since the region of integration D_1 does not include the region $p^2+q^2 \geq 1$, none of the difficulties mentioned in Sec. 1, Part B. are present. The only critical point of significance is the interior stationary point p_s, q_s. The critical points of the phase function on the boundary of D_1 do not contribute because the amplitude function and its first N derivatives all vanish there. Since the region Ω_2 lies entirely within D_1, it then follows immediately from the proof of Theorem 1 of Ref. 11, that

$$u_1(x,y,z)=u_0(x,y,z)+R(x,y,z), \tag{2.18}$$

where

$$R(x,y,z)=O[(kR)^{-N}] \quad \text{as} \quad kR\to\infty$$

with fixed k. It is easy to see from the proof of Theorem 1 in Ref. 11 that $R(x,y,z)=O[(kR)^{-N}]$ uniformly with respect to ξ_1, ξ_2 for all real ξ_1, ξ_2 such that $\delta < (1-\xi_1^2-\xi_2^2)^{1/2} \leq 1$ for any positve constant $\delta < 1$. Hence, only the proof that $u_2(x,y,z)+u_3(x,y,z)=O[(kR)^{-N}]$ uniformly remains.

We can dispose of $u_3(x,y,z)$ easily since m is purely imaginary and $im \leq -[(C_3-\epsilon)^2-1]^{1/2} < 0$ in D_3. For sufficiently large R, there is a positive constant a such that $z > a$. For such R, we have

$$\begin{aligned}|u_3(x,y,z)| &\leq \int\int_{D_3} |U(p,q)|\exp(ikmz)\,dp\,dq \\ &= \int\int_{D_3} |U(p,q)|\exp(ikma)\exp[ikm(z-a)]\,dp\,dq \\ &\leq \exp\{-k(z-a)[(C_3-\epsilon)^2-1]^{1/2}\} \\ &\quad\times \int\int_{D_3} |U(p,q)|\exp(ikma)\,dp\,dq. \end{aligned} \tag{2.19}$$

Since $U(p,q) \in T_N$ and since $a > 0$, it follows from condition (i) in the definition of T_N that the final integral in (2.19) converges to a finite constant M_3 independent of ξ_1, ξ_2, ξ_3. Also since $z=z_0+\xi_3 R$ and $\xi_3 > \delta$, we have

$$|u_3(x,y,z)| \leq M_3 \exp\{-[k(z_0-a)+kR\delta][(C_3-\epsilon)^2-1]^{1/2}\}. \tag{2.20}$$

But the multiplicative constant and the decay constant in the exponential are independent of ξ_1, ξ_2. Hence we have $u_3(x,y,z)=O[(kR)^{-N}]$ uniformly with respect to ξ_1, ξ_2 as $kR\to\infty$ for arbitrary positive N.

Finally, we deal with $u_2(x,y,z)$. First, we make the change of integration variables

$$p=\sin\alpha\cos\beta, \tag{2.21}$$

$$q=\sin\alpha\sin\beta. \tag{2.22}$$

The Jacobian of the transformation is

$$J\left(\frac{p,q}{\alpha,\beta}\right)=\sin\alpha\cos\alpha, \tag{2.23}$$

and we have

$$m=\cos\alpha. \tag{2.24}$$

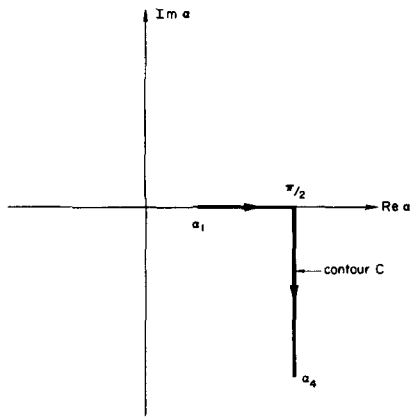

FIG. 2. The contour of integration in Eq. (2.28).

Similarly, we introduce new angles θ, φ such that $0 \leq \theta < \pi/2$, $0 \leq \varphi \leq 2\pi$ defined by

$$\xi_1 = \sin\theta \cos\varphi, \tag{2.25}$$

$$\xi_2 = \sin\theta \sin\varphi. \tag{2.26}$$

These angles are the spherical polar coordinates of the point of intersection on the unit sphere of a line through the origin in x, y, z space, parallel to the direction taken by the point of observation as it recedes from the point (x_0, y_0, z_0). In analogy with (2.24), we have

$$\xi_3 = \cos\theta. \tag{2.27}$$

In terms of these variables, the integral for $u_2(x, y, z)$ in (2.17) becomes

$$u_2(x, y, z) = \int_0^{2\pi} \int_C A(\alpha, \beta) \exp\{ikR[\sin\theta \sin\alpha \cos(\beta - \varphi) + \cos\theta \cos\alpha]\} d\alpha \, d\beta, \tag{2.28}$$

where

$$A(\alpha, \beta) = \nu_2(p, q) V(p, q, m) \exp[ik(px_0 + qy_0 + mz_0)] \sin\alpha, \tag{2.29}$$

with p, q given by (2.21)—(2.22) and $V(p, q, m)$ defined in (1.9). The contour of integration C for the α integral is shown in Fig. 2. The end points of the contour are

$$\alpha_1 = \arcsin(C_1 - \epsilon), \tag{2.30}$$

$$\alpha_4 = \arcsin(C_4 + \epsilon), \tag{2.31}$$

which correspond to the limits of D_2.

The form of the integrand in (2.28) is greatly improved over that in (2.17). The phase function is analytic and the amplitude function is continuous and has partial derivatives up to order N with respect to the variables of integration over the whole integration range. In return for this improvement, we have gained the complication of a complex contour of integration. Moreover, the argument of the exponential is still complex over part of the integration region.

To study the properties of the argument of the exponential in more detail, we define

$$t = \sin\theta \sin\alpha \cos(\beta - \varphi) + \cos\theta \cos\alpha. \tag{2.32}$$

As α varies over the contour C with β fixed, t varies over a simple curve $C(\beta)$ of finite length in the complex plane. Since the partial derivative

$$\left(\frac{\partial t}{\partial \alpha}\right)_\beta = \sin\theta \cos\alpha \cos(\beta - \alpha) - \cos\theta \sin\alpha \tag{2.33}$$

is an analytic function of complex α for all α and β, the partial derivative $(\partial\alpha/\partial t)_\beta$ is an analytic function of α given by

$$\left(\frac{\partial \alpha}{\partial t}\right)_\beta = \left[\left(\frac{\partial t}{\partial \alpha}\right)_\beta\right]^{-1} \tag{2.34}$$

for all α, β such that $(\partial t/\partial\alpha)_\beta \neq 0$. (Here and throughout the paper, we indicate the quantities to be held constant in partial derivatives by placing their symbols as subscripts outside parentheses enclosing the partial derivative). Since $\cos\theta = \xi_3 > \delta$, we have $(\partial t/\partial\alpha)_\beta \neq 0$ when $\cos\alpha = 0$. Hence, the zeros of $(\partial t/\partial\alpha)_\beta$ occur for values of α such that

$$\tan\alpha = \tan\theta \cos(\beta - \varphi). \tag{2.35}$$

But on the part of C on which α is real, we have $\pi/2 \geq \alpha \geq \theta$ so that $\tan\alpha > \tan\theta \cos(\beta - \varphi)$. And on the part of C on which α is complex, $\tan\alpha$ is imaginary whereas $\tan\theta \cos(\beta - \varphi)$ is real. Hence, $(\partial t/\partial\alpha)_\beta$ has no zeros for α on C with $0 \leq \beta \leq 2\pi$. As a result, $(\partial\alpha/\partial t)_\beta$ is an analytic function of complex α given by (2.34) for α on C and $0 \leq \beta \leq 2\pi$. It follows from this analysis, that we can change the variable of integration in the α integral in (2.28) to the variable t defined in (2.32). The result is

$$u_2(x, y, z) = \int_0^{2\pi} \int_{C(\beta)} A[\alpha(t), \beta] \left(\frac{\partial \alpha}{\partial t}\right)_\beta \exp(ikRt) \, dt \, d\beta. \tag{2.36}$$

Because of the prescribed properties of $\nu_2(p, q)$ and $V(p, q, m)$, the quantity $A(\alpha, \beta)$ has N continuous partial derivatives with respect to α with constant β taken along the complex curve C for all α on C and all $0 \leq \beta \leq 2\pi$. (Note that the derivatives must be taken along the curve C since $\nu_2(p, q)$ and $V(p, q, m)$ are defined only for real p, q. Varying α along C corresponds to varying p, q along the real axis.) Since $(\partial\alpha/\partial t)_\beta$ is an analytic function of α on C, the product $A(\alpha, \beta)(\partial\alpha/\partial t)_\beta$ also has N continuous partial derivatives with respect to α with constant β taken along C. Finally, we note that differentiation with respect to t along the curve $C(\beta)$ with constant β, is equivalent to differentiating with respect to α along C with constant β and multiplying by $(\partial\alpha/\partial t)_\beta$. Hence, it follows from the analyticity of $(\partial\alpha/\partial t)_\beta$ with respect to α that the quantity $A(\alpha, \beta)(\partial\alpha/\partial t)_\beta$ has N continuous partial derivatives with respect to t taken along $C(\beta)$ with constant β for all t on $C(\beta)$ and all $0 \leq \beta \leq 2\pi$. We now integrate the t integral in (2.36),

$$I(kR, \beta) = \int_{C(\beta)} A[\alpha(t), \beta] \left(\frac{\partial \alpha}{\partial t}\right)_\beta \exp(ikRt) \, dt, \tag{2.37}$$

by parts N times by integrating $\exp(ikRt)$ each time and differentiating the rest to obtain[7]

$$I(kR, \beta) = L_N(t_4) - L_N(t_1) + R_N(kR, \beta), \tag{2.38}$$

where

$$L_N(t_j)=\sum_{n=0}^{N-1} i^{n-1}\left\{\frac{\partial^n}{\partial t^n}\left[A(\alpha,\beta)\left(\frac{\partial\alpha}{\partial t}\right)_\beta\right]\right\}_\beta\Bigg|_{t=t_j}\frac{\exp(ikRt_j)}{(kR)^{n+1}}, \tag{2.39}$$

$$R_N(kR,\beta)=(-ikR)^{-N}\int_{C(\beta)}\left\{\frac{\partial^N}{\partial t^N}\left[A(\alpha,\beta)\left(\frac{\partial\alpha}{\partial t}\right)_\beta\right]\right\}_\beta \times\exp(ikRt)\,dt, \tag{2.40}$$

$$t_j=\sin\theta\sin\alpha_j\cos(\beta-\varphi)+\cos\theta\cos\alpha_j, \tag{2.41}$$

with α_1, α_4 given by (2.30), (2.31). Due to the properties of $\nu_2(p,q)$, $A(\alpha,\beta)$ and all N of its partial derivatives with respect to α vanish at the endpoints t_1, t_4 of $C(\beta)$. Hence, the derivatives of $A(\alpha,\beta)$ with respect to t taken along $C(\beta)$ also vanish at t_1, t_4, so we have

$$L_N(t_1)=L_N(t_4)=0. \tag{2.42}$$

Since the integrand in (2.40) is continuous and since the contour $C(\beta)$ is of finite length for all θ, φ, and β, the integral is bounded by some positive constant M_2 independent of θ, φ, β. Hence

$$|R_N(kR,\beta)|\le M_2(kR)^{-N} \tag{2.43}$$

and consequently

$$|u_2(x,y,z)|\le 2\pi M_2(kR)^{-N}. \tag{2.44}$$

Thus, we have shown that $u_2(x,y,z)=O[(kR)^{-N}]$ uniformly with respect to ξ_1, ξ_2 as $kR\to\infty$ with fixed k.

Collecting our results, we find that the three terms in (2.16) that add to give $u(x,y,z)$ have the following properties:

$$u_1(x,y,z)=u_0(x,y,z)+O[(kR)^{-N}], \tag{2.45}$$

$$u_2(x,y,z)=O[(kR)^{-N}], \tag{2.46}$$

$$u_3(x,y,z)=O[(kR)^{-N}], \tag{2.47}$$

uniformly with respect to ξ_1, ξ_2 as $kR\to\infty$ with fixed k. Hence, the theorem is proved.

B. Asymptotic approximation of $u(x, y, z)$

It follows from Sec. 2, Part A that any terms of order lower than $(kR)^{-N}$ in the asymptotic behavior of $u(x,y,z)$ must be contributed by $u_0(x,y,z)$. The integral representation of $u_0(x,y,z)$ in (2.5) can be put in the form of $I(kR)$ in (1.15) with phase function

$$f(p,q)=\xi_1 p+\xi_2 q+\xi_3 m \tag{2.48}$$

and amplitude function

$$g(p,q)=\nu(p,q)U(p,q)\exp[ik(px_0+qy_0+mz_0)]. \tag{2.49}$$

Since $\nu(p,q)=0$ for p,q not in Ω_2, we can take the domain D of integration to be a region within the interior of D_H containing Ω_2 within its interior. Then $f(p,q)$ in (2.48) is real and infinitely differentiable in D, and $g(p,q)$ has continuous, bounded partial derivatives up to order N in D. Hence, the integral satisfies all of the conditions required for the application of Ref. 9 to obtain the asymptotic behavior of $u_0(x,y,z)$. Since there is only one critical point in D (the interior stationary point $p=\xi_1$, $q=\xi_2$) and since the integrand and its derivatives all vanish on the boundary of D, it follows immediately from Ref. 9 that

$$u_0(x,y,z)=\frac{\exp(ikR)}{kR}\sum_{n=0}^{N/2}\frac{B_n(\theta,\varphi)}{(kR)^n}+O[(kR)^{-1-N/6}], \tag{2.50}$$

where the coefficients $B_n(\theta,\varphi)$ are independent of R. (It may be helpful to recall at this point that N is specified in the definition of T_N to be an *even* positive integer.)

The estimate of the order of the remainder term in (2.50) has been improved by Stamnes and Sherman.[18] They show that the remainder term is $o[(kR)^{-N/2}]$. It follows then from the results of Sec. 2, Part A. combined with Eq. (2.50) and the results of Ref. 18 that the asymptotic behavior of the total angular spectrum integral $u(x,y,z)$ is given by

$$u(x,y,z)=[\exp(ikR)/kR)]\sum_{n=0}^{N/2-1}B_n(\theta,\varphi)/(kR)^n +o[(kR)^{-N/2}] \tag{2.51}$$

as $kR\to\infty$ with fixed k, ξ_1, ξ_2 for $U(p,q)\in T_N$.

Although Braun[9] gives formulas that completely determine the coefficients $B_n(\theta,\varphi)$ in terms of the amplitude and phase functions of the integral, application of these formulas to obtain explicit relations between $B_n(\theta,\varphi)$ and $U(p,q)$ is very involved for all $n>0$. In the hope that they may prove simpler to apply, we can turn to the quite different expressions for $B_n(\theta,\varphi)$ provided by Jones and Kline.[10] Although these expressions were derived under the assumption that $g(p,q)$ is analytic in a neighborhood of $p=\xi_1$, $q=\xi_2$, they must yield the same results as those in Ref. 9 since (a) the case treated in Ref. 10 is included in the analysis of Ref. 9 with $N=\infty$, (b) the functional dependence of $B_n(\theta,\varphi)$ on $U(p,q)$ as given in Ref. 9 is independent of N, and (c) the asymptotic power series of a function is unique. Unfortunately, the formulas of Jones and Kline also are cumbersome to apply for $n>0$.

The formulas for the first term however, are simple. Applying Sec. 5.1 of Ref. 10 we obtain the well-known result

$$B_0(\theta,\varphi)=-2\pi i\exp[ik(\xi_1x_0+\xi_2y_0+\xi_3z_0)] \times V(\xi_1,\xi_2,\xi_3), \tag{2.52}$$

with $V(p,q,m)$ given by (1.9) and with ξ_1, ξ_2, ξ_3 given by (1.6)—(1.8) and related to θ, φ by (2.25)—(2.27). Application of Ref. 10 to find the second-order term is much more difficult. After a very lengthy but straightforward calculation, we find the relatively simple result

$$B_1(\theta,\varphi)=(i/2)L^2B_0(\theta,\varphi), \tag{2.53}$$

where L^2 is the differential operator defined by

$$L^2=-(1/\sin\theta)\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial}{\partial\theta}\right)-(1/\sin^2\theta)\frac{\partial^2}{\partial\varphi^2}. \tag{2.54}$$

Hence, if $U(p,q)\in T_N$ with $N\ge 6$, we have

$$u(x,y,z)=-2\pi i[\exp(ikR)/kR][1+(i/2kR)L^2] \times\exp[ik(\xi_1x_0+\xi_2y_0+\xi_3z_0)]V(\xi_1,\xi_2,\xi_3) +O[(kR)^{-3}] \tag{2.55}$$

as $kR\to\infty$ with fixed k, ξ_1, ξ_2.

The relation between B_0 and B_1 given in (2.53), along with a recursion formula for higher order coefficients, can be obtained much more simply by making use of the

fact that $u(x, y, z)$ satisfies the Helmholtz equation (1.4) for $z > 0$. Sherman[19] has shown that if a solution $u(x, y, z)$ of (1.4) has an asymptotic expansion of the form in (2.51) for arbitrarily large N and if the asymptotic expansions of the partial derivatives of $u(x, y, z)$ with respect to R, θ, φ up to order 2 can be obtained by differentiating (2.51) term by term, then the coefficients $B_n(\theta, \varphi)$ satisfy the recursion formula

$$(n+1)B_{n+1}(\theta, \varphi) = (i/2)[L^2 - n(n+1)]B_n(\theta, \varphi). \quad (2.56)$$

It can be shown that the partial derivatives of $u(x, y, z)$ given in (1.1) satisfy the above requirements. Hence, the coefficients $B_n(\theta, \varphi)$ satisfy (2.56) if $U(p, q) \in T_\infty$. As argued before, this means that (2.56) must hold for $U(p, q) \in T_N$ with N finite as well, since the functional dependence of the coefficients on $U(p, q)$ is independent of N. Hence, (2.56) provides a straightforward method for obtaining higher order terms in the asymptotic approximation of $u(x, y, z)$ from a knowledge of the first term.

3. APPROXIMATION VALID ON THE PLANE $z = z_0$

In this section, we obtain an asymptotic approximation of $u(x, y, z)$ valid when the point of observation recedes towards infinity in a fixed direction perpendicular to the z axis through a point (x_0, y_0, z_0) in the half-space $z > 0$. This case is excluded in Sec. 2 because the analysis therein cannot be extended to include $\xi_3 = 0$. The problem in this section is simpler in principle than that of Sec. 2 because an extension of the MSP is not required. But the calculation of the asymptotic approximation is much more involved in this case because several critical points must be considered. We again take $U(p, q) \in T_N$, and we let $z = z_0 > 0$.

When we set $z = z_0$ in (1.2), the resulting integral is of the form in (1.15) with

$$f(p, q) = \xi_1 p + \xi_2 q, \quad (3.1)$$

$$g(p, q) = U(p, q) \exp[ik(px_0 + qy_0 + mz_0)], \quad (3.2)$$

$$D = D_J \quad (J = H, I). \quad (3.3)$$

Hence, we have a phase function $f(p, q)$ that is real and analytic in both D_H and D_I. Unfortunately, the amplitude function $g(p, q)$ need not have any partial derivatives on the circle $p^2 + q^2 = 1$ in general. Therefore, we isolate that circle using a neutralizer. Let $\nu'(p, q)$ be a real, continuous function with continuous partial derivatives of all orders for all p, q. Moreover, let $\nu'(p, q)$ satisfy

$$0 \leq \nu'(p, q) \leq 1, \quad (3.4)$$

$$\nu'(p, q) = 1 \quad \text{for} \quad 1 - \epsilon \leq p^2 + q^2 \leq 1 + \epsilon, \quad (3.5)$$

$$\nu'(p, q) = 0 \quad \text{for} \quad p^2 + q^2 \leq 1 - 2\epsilon \quad \text{and} \quad p^2 + q^2 \geq 1 + 2\epsilon, \quad (3.6)$$

where ϵ is a positive constant less than $\frac{1}{3}$. Then, we can write the integral in (1.2) in the form (with $J = H, I$)

$$u_J(x, y, z_0) = u_{J1}(x, y, z_0) + u_{J2}(x, y, z_0), \quad (3.7)$$

where

$$u_{J1}(x, y, z_0) = \int\int_{D_{J1}} \nu'(p, q)\, U(p, q) \times \exp[ik(px + qy + mz_0)]\, dp\, dq, \quad (3.8)$$

$$u_{J2}(x, y, z_0) = \int\int_{D_{J2}} [1 - \nu'(p, q)]U(p, q) \times \exp[ik(px + qy + mz_0)]\, dp\, dq, \quad (3.9)$$

D_{H1} is the region $1 - 3\epsilon \leq p^2 + q^2 \leq 1$,

D_{H2} is the region $p^2 + q^2 \leq 1 - \epsilon/2$,

D_{I1} is the region $1 \leq p^2 + q^2 \leq 1 + 3\epsilon$,

D_{I2} is the region $p^2 + q^2 \geq 1 + \epsilon/2$.

The various regions of interest in the p, q plane are shown in Fig. 3.

The integrand in the expression for u_{H2} satisfies all conditions required for application of Theorem 1 of Ref. 11. Since there are no critical points of the integrand within D_{H2}, and since the amplitude function and all N of its partial derivatives vanish on the boundary of D_{H2}, we have

$$u_{H2} = O[(kR)^{-N}] \quad (3.10)$$

as $kR \to \infty$ with fixed k.

The integrand in the expression for u_{I2} also satisfies the same conditions but the region D_{I2} is infinite in extent whereas Theorem 1 of Ref. 11 was proved only for a finite region of integration. The theorem can be extended without difficulty, however, to include our case. We simply consider the same integral, but with region of integration $1 + \epsilon/2 \leq p^2 + q^2 \leq K$, and then follow the original proof. The main new feature is the presence of new terms that arise because the amplitude function and its partial derivatives do not vanish on the boundary $p^2 + q^2 = K$. These terms do vanish, however, when we take the limit $K \to \infty$, because the amplitude and all of its derivatives contain the factor $\exp[-(K-1)^{1/2} z_0]$ with $z_0 > 0$. (Note that the case $z_0 = 0$ can be treated only by placing additional restrictions on the behavior of $U(p, q)$ and its derivatives as $p^2 + q^2 \to \infty$.) The other new feature is that the remainder integral after N integration by parts is over an infinite region. But all that is required in the proof is that the integral is convergent. Convergence is guaranteed, however, by the presence of a factor $\exp[-(p^2 + q^2 - 1)^{1/2} z_0]$ in the integrand. Hence, the result of the theorem applies without change, and we have

$$u_{I2}(x, y, z_0) = O[(kR)^{-N}] \quad (3.11)$$

as $kR \to \infty$ with fixed k.

To deal with $u_{H1}(x, y, z)$, we make the change of variables of integration given in (2.21)—(2.22) and use the notation of (2.25)—(2.26) to obtain

$$u_{H1}(x, y, z_0) = \int_{\beta_0}^{2\pi + \beta_0} \int_{\alpha_1'}^{\pi/2} A'(\alpha, \beta) \exp[ikR \sin\alpha \cos(\beta - \varphi)]\, d\alpha\, d\beta, \quad (3.12)$$

where $\alpha_1' = \arcsin(1 - 3\epsilon)^{1/2} = \arccos(3\epsilon)^{1/2}$ and $A'(\alpha, \beta)$

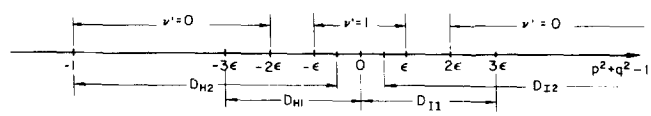

FIG. 3. Illustration of the notation used in Eqs. (3.4)—(3.9).

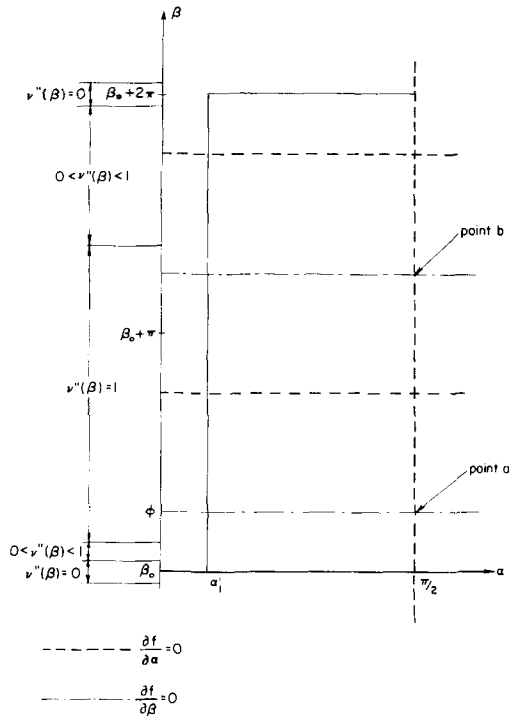

FIG. 4. The region of integration and the location of the critical points in Eq. (3.12).

is the same as $A(\alpha, \beta)$ defined by (2.29) except that $\nu_2(p, q)$ is replaced by $\nu'(p, q)$. Since the constant β_0 is arbitrary, we choose it for later convenience to be $\beta_0 = \varphi - \pi/4$. The integral is now in the form in (1.15) (with integration variables α, β instead of p, q) with phase function

$$f(\alpha, \beta) = \sin\alpha \cos(\beta - \varphi) \tag{3.13}$$

and amplitude function

$$g(\alpha, \beta) = A'(\alpha, \beta). \tag{3.14}$$

Since $f(\alpha, \beta)$ and $g(\alpha, \beta)$ satisfy all of the requirements of the MSP over the region of integration, we are now in a position to approximate $u_{H1}(x, y, z)$.

The region of integration and the location of the critical points of the integral in (3.12) are shown in Fig. 4. All of the critical points lie on the boundary of the integration region. The critical points on the boundary $\alpha = \alpha_1'$ do not contribute to the asymptotic behavior of $u_{H1}(x, y, z)$ because the amplitude function and all N of its partial derivatives vanish there. The remaining critical points are two corners of the boundary ($\alpha = \pi/2$, $\beta = \beta_0$ and $\alpha = \pi/2$, $\beta = 2\pi + \beta_0$) and two stationary points ($\alpha = \pi/2$, $\beta = \varphi$ and $\alpha = \pi/2$, $\beta = \varphi + \pi$). Because of our choice of β_0, neither of the stationary points coincide with a boundary corner.

Consider first the corner critical points. These are an artificial creation of the change of variables of integration. Since they can be shifted freely along the line $\alpha = \pi/2$ in the α, β plane by different choices of β_0, it is reasonable to expect that taken together, the two corner points do not contribute to the asymptotic behavior of $u(x, y, z)$. This expectation can be verified by the following argument. We construct a neutralizer $\nu''(\beta)$, a real continuous function with continuous derivatives of all orders, that is periodic with period 2π, is equal to 1 for $\varphi - \pi/8 \le \beta \le \varphi + 9\pi/8$ and is zero in a neighborhood of $\beta = \beta_0$. Then we write $u_{H1}(x, y, z_0)$ as the sum of two integrals, each the same as the one in (3.12) except that the first integral contains an extra factor $\nu''(\beta)$ whereas the second integral contains an extra factor $1 - \nu''(\beta)$. The only critical points of importance in the first integral are the stationary points already described because the integrand vanishes at the corner points. Moreover, since $\nu''(\beta) = 1$ in a neighborhood of the stationary points, the asymptotic behavior of this integral is precisely the same as the contribution of those stationary points to the asymptotic behavior of $u_{H1}(x, y, z_0)$. The only critical points of importance in the second integral are the corner points since $1 - \nu''(\beta)$ vanishes at the stationary points. The integrand can be made to vanish at the corner points, however, by changing the region of integration to be $\varphi - 9\pi/8 \le \beta \le \varphi + 7\pi/8$. This is possible since the integrand is a periodic function of β with period 2π. Hence, the integrand vanishes in the vicinity of all critical points, and it follows from Theorem 1 of Ref. 11 that the second integral is $O[(kR)^{-N}]$. We conclude that the asymptotic behavior of $u_{H1}(x, y, z_0)$ of order lower than $(kR)^{-N}$ is given solely by the contributions of the two stationary points.

Now consider the stationary points. Let point a be the point $\alpha = \pi/2$, $\beta = \varphi$ and point b be the point $\alpha = \pi/2$, $\beta = \varphi + \pi$. Further, let $u_{H1}^{(a)}(x, y, z_0)$ and $u_{H1}^{(b)}(x, y, z_0)$ be the contributions to the asymptotic behavior of $u_{H1}(x, y, z_0)$ due to points a and b respectively. Then, we have

$$u_{H1}(x, y, z_0) = u_{H1}^{(a)}(x, y, z_0) + u_{H1}^{(b)}(x, y, z_0) + O[(kR)^{-N}]. \tag{3.15}$$

To obtain the asymptotic approximation of $u_{H1}^{(a)}$ and $u_{H1}^{(b)}$ in (3.15), we need results that apply to the case when the amplitude function has only a finite number of derivatives. The boundary stationary point of the elliptic type has been treated by Braun and the result for the order of the remainder term is similar to that given in (2.50). This estimate has been improved by Stamnes and Sherman.[18] The remainder term pertaining to a boundary stationary point of the hyperbolic type is not yet available in the literature, but is shown by Stamnes and Sherman[18] to be $o(k^{-N/4})$. Also, the form for the asymptotic sequence for the hyperbolic stationary point is the same as that for the elliptic point. Thus we have

$$u_{H1}^{(j)}(x, y, z_0) = u_{H1}^{(j1)}(x, y, z_0) + u_{H1}^{(j2)}(x, y, z_0), \tag{3.16}$$

where

$$u_{H1}^{(j1)}(x, y, z_0) = [\exp(\pm ikR)/kR] \sum_{n=0}^{N/2} B_{Hn}^{(j1)}(\varphi)/(kR)^n + o[(kR)^{-N/4}], \tag{3.17}$$

$$u_{H1}^{(j2)}(x, y, z_0) = [\exp(\pm ikR)/(kR)^{3/2}] \sum_{n=0}^{N/2-1} B_{Hn}^{(j2)}(\varphi)/kR)^n + o[(kR)^{-N/4}], \tag{3.18}$$

with $B_{Hn}^{(j1)}(\varphi)$ and $B_{Hn}^{(j2)}(\varphi)$ independent of R. In (3.17) and (3.18), the upper sign in the exponential applies when $j = a$ and the lower sign applies when $j = b$.

To obtain the asymptotic expansion of $u_{I1}(x, y, z_0)$, we change the variables of integration in (3.8) to new variables μ, λ defined by

$$p = \cosh\mu \cos\lambda, \tag{3.19}$$

$$q = \cosh\mu \sin\lambda. \tag{3.20}$$

The Jacobian of the transformation is

$$J\left(\frac{p, q}{\mu, \lambda}\right) = \sinh\mu \cosh\mu, \tag{3.21}$$

and we have

$$m = +i \sinh\mu. \tag{3.22}$$

With these new variables, we have

$$u_{I1}(x, y, z_0) = \int_{\beta_0}^{\beta_0+2\pi} \int_0^{\mu_1} A''(\mu, \lambda) \exp[ikR \cosh\mu \cos(\lambda - \varphi)] \times d\mu\, d\lambda, \tag{3.23}$$

where

$$\mu_1 = \operatorname{arcsinh}(3\epsilon)^{1/2} = \operatorname{arccosh}(1 + 3\epsilon)^{1/2},$$

$$A''(\mu, \lambda) = -i\nu'(p, q) V(p, q, m) \exp[ik(px_0 + qy_0 + mz_0)] \cosh\mu, \tag{3.24}$$

with p, q given by (3.19)—(3.20). The integral is now in the form of (1.15) with phase and amplitude functions

$$f(\mu, \lambda) = \cosh\mu \cos(\lambda - \varphi), \tag{3.25}$$

$$g(\mu, \lambda) = A''(\mu, \lambda). \tag{3.26}$$

As with the integral for $u_{H1}(x, y, z_0)$, all of the critical points lie on the boundary. For the same reasons as in the previous case, the critical points on the boundary $\mu = \mu_1$ and the corners at $\mu = 0$, $\lambda = 0$ and $\mu = 0$, $\lambda = 2\pi$ do not contribute. The asymptotic behavior of $u_{I1}(x, y, z_0)$ is determined completely by stationary point a at $\mu = 0$, $\lambda = \varphi$ and stationary point b at $\mu = 0$, $\lambda = \varphi + \pi$. As before, we write

$$u_{I1}(x, y, z_0) = u_{I1}^{(a)}(x, y, z_0) + u_{I1}^{(b)}(x, y, z_0) + O[(kR)^{-N}] \tag{3.27}$$

where $u_{I1}^{(j)}(x, y, z_0)$ is the contribution due to point j. We then again apply Ref. 9 and the results of Stamnes and Sherman[18] and find that the asymptotic approximations of $u_{I1}^{(a)}(x, y, z_0)$ and $u_{I1}^{(b)}(x, y, z_0)$ are of the same form as those of $u_{H1}^{(a)}(x, y, z_0)$ and $u_{H1}^{(b)}(x, y, z_0)$ given in (3.16)—(3.18) but with coefficients $B_{In}^{(j1)}(\varphi)$ and $B_{In}^{(j2)}(\varphi)$ in place of $B_{Hn}^{(j1)}(\varphi)$ and $B_{Hn}^{(j2)}(\varphi)$ respectively.

Comparison of the explicit values of the various coefficients remains. The expressions for the coefficients given in Ref. 10 can be used for that purpose for the same reasons as given in Sec. 2, Part B. to justify the application of the results of Ref. 10 there. We have found, however, that the formula (47) of Ref. 10 which applies in the present case is incorrect. In Appendix A, we derive the correct formula by applying the procedure outlined in Ref. 10, and in Appendix B, we use the result to obtain expressions for the coefficients. The final results are

$$B_{In}^{(a1)}(\varphi) = B_{Hn}^{(a1)}(\varphi), \tag{3.28}$$

$$B_{In}^{(a2)}(\varphi) = -B_{Hn}^{(a2)}(\varphi), \tag{3.29}$$

$$B_{In}^{(b1)}(\varphi) = -B_{Hn}^{(b1)}(\varphi), \tag{3.30}$$

$$B_{In}^{(b2)}(\varphi) = -B_{Hn}^{(b1)}(\varphi). \tag{3.31}$$

It follows from (3.30)—(3.31), that the contribution of point b to $u_{I1}(x, y, z_0)$ is equal in magnitude and opposite in sign to the contribution of point b to $u_{H1}(x, y, z_0)$. Hence, the point b does not contribute to the asymptotic behavior [of order lower than $(kR)^{-N/4}$] of the total integral $u_{H1}(x, y, z_0) + u_{I1}(x, y, z_0)$. Similarly, (3.29) shows that the contribution of point a to the series involving inverse half-powers of (kR) in the expansion of $u_{I1}(x, y, z_0)$ is equal in magnitude and opposite in sign to the same contribution of point a to $u_{H1}(x, y, z_0)$. Hence the expansion of the total integral $u_{H1}(x, y, z_0) + u_{I1}(x, y, z)$ does not include half-powers of kR of order lower than $(kR)^{-N/4}$. Finally, it follows from (3.28) that $u_{H1}(x, y, z_0)$ and $u_{I1}(x, y, z_0)$ contribute equally to the remaining terms involving only integral inverse powers of kR that make up the expansion of the total integral. Hence we have

$$u_{H1}(x, y, z_0) + u_{I1}(x, y, z_0) = 2u_{H1}^{(a1)}(x, y, z_0). \tag{3.32}$$

Therefore, in view of (1.1), (3.7), (3.10), (3.11), (3.17) and (3.32) the asymptotic behavior of $u(x, y, z_0)$ is given by

$$u(x, y, z_0) = [\exp(ikR)/kR] \sum_{n=0}^{N/2} 2B_{Hn}^{(a1)}(\varphi)/(kR)^n + o[(kR)^{-N/4}] \tag{3.33}$$

as $kR \to \infty$ with constant k and $z = z_0$.

Equation (3.33) provides the desired asymptotic approximation of $u(x, y, z_0)$. Only the comparison of the result with the result of Sec. 2, Part B remains. In Appendix C, we compare the coefficients $B_{Hn}^{(a1)}(\varphi)$ appearing in (3.33) with the coefficients $B_n(\theta, \varphi)$ appearing in (2.51); the result is

$$B_{Hn}^{(a1)}(\varphi) = \tfrac{1}{2} \lim_{\theta \to \pi/2} B_n(\theta, \varphi). \tag{3.34}$$

Hence, to order lower than $(kR)^{-N/4}$, the series in (3.33) is equivalent to the series in (2.51) with θ set equal to $\pi/2$. Consequently, we can combine the results of this section and Sec. 2 to write

$$u(x, y, z) = [\exp(ikR)/kR] \sum_{n=0}^{N/2} B_n(\theta, \varphi)/(kR)^n + o[(kR)^{-N/4}] \tag{3.35}$$

as $kR \to \infty$ with constant k for $z_0 > 0$, $0 \le \varphi \le 2\pi$, and $0 \le \theta \le \pi/2$. The asymptotic expansion (3.35) is therefore uniform with respect to the angles θ and φ in the domain $0 \le \theta \le \pi/2$ and $0 \le \varphi \le 2\pi$. The coefficients $B_n(\theta, \varphi)$ for $n = 0$ and $n = 1$ are given in (2.55).

Finally, we note that in the important special case when $V(p, q, m)$ is given by (1.13), the first term in the series in (3.35) vanishes for $\theta = \pi/2$. The dominant term in the expansion is then the term with $n = 1$ in (3.35). Applying the expression in (2.55) for that term, we find that for $N \ge 12$,

$$u(x, y, z_0) = -2\pi \exp[ik(\xi_1 x_0 + \xi_2 y_0)] \times [\exp(ikR)/(kR)^2] \left[ikz_0 W(\xi_1, \xi_2, \xi_3) + \frac{\partial W(\xi_1, \xi_2, s)}{\partial s}\bigg|_{s=\xi_3} \right] + O[(kR)^{-3}] \tag{3.36}$$

as $kR \to \infty$ with constant k for $z_0 > 0$, $0 \le \varphi \le 2\pi$, $\theta = \pi/2$.

4. APPROXIMATIONS OF u_H (x, y, z) AND u_I (x, y, z)

In this section, we obtain the dominant term in the asymptotic behavior of $u_H(x, y, z)$ and $u_I(x, y, z)$. The behavior is found to be quite different in the three different cases $0 < \xi_3 < 1$, $\xi_3 = 0$, and $\xi_3 = 1$.

A. Approximations valid over the hemisphere $0 < \xi_3 < 1$

To approximate $u_H(x, y, z)$, we apply the notation and definitions used in (3.4)—(3.9) with z_0 replaced by z. The only critical point of the integral $u_{H2}(x, y, z)$, where the integrand is nonzero, is the interior stationary point $p_s = \xi_1$, $q_s = \xi_2$. Hence, the asymptotic behavior of $u_{H2}(x, y, z)$ is identical to that of $u(x, y, z)$ as given in (2.51) and (2.55). It follows then from (1.1) that the asymptotic behavior of $u_I(x, y, z)$ is identical to that of $-u_{H1}(x, y, z)$.

The integral $u_{H1}(x, y, z)$ is dealt with in the same way as was $u_{H1}(x, y, z_0)$ in Sec. 3. The change of variables of integration given in (2.21)—(2.22) is made, yielding an integral identical to the one in (3.12) except that the phase function $f(\alpha, \beta)$ [in (3.13)] is replaced by

$$f(\alpha, \beta) = \sin\theta \sin\alpha \cos(\beta - \varphi) + \cos\theta \cos\alpha. \tag{4.1}$$

This change in phase function does not affect the location of the critical points; they are still as shown in Fig. 4. Again, the only critical points that contribute to the asymptotic behavior are the points a and b. In this case, however, the critical points a and b are no longer ordinary stationary points where both $\partial f/\partial\alpha$ and $\partial f/\partial\beta$ vanish. Instead, they are nonstationary points on the boundary where the lines of constant $f(\alpha, \beta)$ are tangent to the boundary (i.e., points on the boundary $\alpha = \pi/2$ where $\partial f/\partial\beta = 0$).

Such critical points are treated in detail in Sec. 4.1 of Ref. 10 for the case when the amplitude function $A'(\alpha, \beta)$ is an analytic function of α, β. The case when $A'(\alpha, \beta)$ has only a finite number of partial derivatives is discussed in Ref. 9. Braun[9] gives the form of the asymptotic sequence, but does not determine the number of terms that can be obtained if $A'(\alpha, \beta)$ has partial derivatives up to order N or give an estimate of the order of the remainder term. It can be shown, however[18] that the first term in the series plus a remainder term that is $O[(kR)^{-2}]$ can be obtained if $N \geq 8$.

As in the previous sections, the expression for the coefficient of the first term can be obtained by applying the results in Ref. 10. The result for $u_{H1}(x, y, z)$ is

$$\begin{aligned} u_{H1}(x, y, z) = {} & (2\pi/\sin\theta)^{1/2} [\exp(i\pi/4)/(kR)^{3/2} \cos\theta] \\ & \times [V(\xi_1', \xi_2', 0) \exp[ik(x_0\xi_1' + y_0\xi_2')] \exp(ikR \sin\theta) \\ & + iV(-\xi_1', -\xi_2', 0) \exp[-ik(x_0\xi_1' + y_0\xi_2')] \\ & \times \exp(-ikR \sin\theta)] + O[(kR)^{-2}], \end{aligned} \tag{4.2}$$

where

$$\xi_1' = \xi_1/\sin\theta = \cos\varphi, \tag{4.3}$$

$$\xi_2' = \xi_2/\sin\theta = \sin\varphi. \tag{4.4}$$

Hence, the desired asymptotic approximations for $u_H(x, y, z)$ and $u_I(x, y, z)$ are for $N \geq 8$

$$\begin{aligned} u_H(x, y, z) = {} & -2\pi i [\exp(ikR)/kR] V(\xi_1, \xi_2, \xi_3) \\ & \times \exp[ik(\xi_1 x_0 + \xi_2 y_0 + \xi_3 z_0)] + u_{H1}(x, y, z) + O[(kR)^{-2}], \end{aligned} \tag{4.5}$$

$$u_I(x, y, z) = -u_{H1}(x, y, z) + O[(kR)^{-2}], \tag{4.6}$$

as $kR \to \infty$ with fixed ξ_1, ξ_2, k and with $u_{H1}(x, y, z)$ given by (4.2).

The results show that $u_I(x, y, z)$ is of higher order in $(kR)^{-1}$ than is $u_H(x, y, z)$. Hence, we have provided a rigorous justification for neglecting $u_I(x, y, z)$ compared to $u_H(x, y, z)$ for large kR when $0 < \xi_3 < 1$. The results show also that $u_I(x, y, z)$ is of even higher order in $(kR)^{-1}$ compared to $u_H(x, y, z)$ when $V(p, q, 0)$ is zero as is the case when $V(p, q, m)$ is of the form given in (1.13).

B. Approximations valid on the plane $z = z_o$ ($\xi_3 = 0$)

The asymptotic approximations required for this case were obtained in Sec. 3 as part of the analysis of $u(x, y, z_0)$. In this section, we present the dominant terms explicitly. The asymptotic approximation of $u_H(x, y, z_0)$ is given in (3.7), (3.10), and (3.15)—(3.18) with the coefficients in the series given in Appendix B, (B10)—(B11). Combining these expressions and evaluating the first term, we find for $N \geq 8$

$$\begin{aligned} u_H(x, y, z_0) = {} & -(i\pi/kR)\{\exp(ikR) V(\xi_1, \xi_2, 0) \exp[ik(x_0\xi_1 + y_0\xi_2)] \\ & - \exp(-ikR) V(-\xi_1, -\xi_2, 0) \exp[-ik(x_0\xi_1 + y_0\xi_2)]\} \\ & + O[(kR)^{-3/2}]. \end{aligned} \tag{4.7}$$

Similarly, combining (3.7), (3.11), (3.27), (3.28), and (3.30), we find for $N \geq 8$

$$\begin{aligned} u_I(x, y, z_0) = {} & -(i\pi/kR)\{\exp(ikR) V(\xi_1, \xi_2, 0) \exp[ik(x_0\xi_1 + y_0\xi_2)] \\ & + \exp(-ikR) V(-\xi_1, -\xi_2, 0) \exp[-ik(x_0\xi_1 + y_0\xi_2)]\} \\ & + O[(kR)^{-3/2}] \end{aligned} \tag{4.8}$$

as $kR \to \infty$ with fixed ξ_1, ξ_2, k. We see that in this case, $u_H(x, y, z_0)$ and $u_I(x, y, z_0)$ are of the same order in $(kR)^{-1}$, so that $u_I(x, y, z_0)$ cannot be neglected compared to $u_H(x, y, z_0)$.

In the special case when $V(p, q, m)$ is of the form given in (1.13), the coefficients given in (4.7)—(4.8) vanish. Hence, to obtain the dominant term in that case, we must calculate the first-order term given in (3.18). The result is for $N \geq 10$

$$\begin{aligned} u_H(x, y, z_0) = {} & -u_I(x, y, z_0) \\ = {} & [\sqrt{2\pi}/(kR)^{3/2}] \exp(i\pi/4)\{\exp(ikR) \\ & \times \exp[ik(x_0\xi_1 + y_0\xi_2)] W(\xi_1, \xi_2, 0) \\ & + i \exp(-ikR) \exp[-ik(x_0\xi_1 + y_0\xi_2)] \\ & \times W(-\xi_1, -\xi_2, 0)\} + O[(kR)^{-2}] \end{aligned} \tag{4.9}$$

as $kR \to \infty$ with fixed k, ξ_1, ξ_2. Again $u_H(x, y, z_0)$ and $u_I(x, y, z_0)$ are of the same order in $(kR)^{-1}$. Also, we see that they are of lower order than $u(x, y, z_0)$ which is given in (3.36).

C. Approximations valid on the line $x = x_0$, $y = y_0$ ($\xi_3 = 1$)

In this case, $u_H(x, y, z)$ can be written

$$u_H(x_0, y_0, z) = \int_0^{2\pi}\int_0^{\pi/2} A(\alpha, \beta) \exp[ik(z - z_0)\cos\alpha]\, d\alpha\, d\beta, \tag{4.10}$$

where

$$A(\alpha, \beta) = V(p, q, m) \exp[ik(px_0 + qy_0 + mz_0)] \sin\alpha, \tag{4.11}$$

with p, q, m related to α, β by (2.21), (2.22), and (2.24). Since the phase function in (4.10) does not involve β, the integral can be treated most simply by considering it to be a single integral of the form

$$u_H(x_0, y_0, z) = \int_0^{\pi/2} A(\alpha) \exp(ikR\cos\alpha)\, d\alpha, \tag{4.12}$$

where

$$A(\alpha) = \int_0^{2\pi} A(\alpha, \beta)\, d\beta, \tag{4.13}$$

and apply the method of stationary phase for single integrals as given by Erdélyi.[7] The contributions to the asymptotic behavior come from the end points of integration $\alpha = 0$ and $\alpha = \pi/2$. The result for the dominant term is (for $N \geq 4$)

$$u_H(x_0, y_0, z) = -2\pi i[\exp(ikR)/kR]\, V(0, 0, 1) \exp(ikz_0) + (i/kR)A(\pi/2) + O[(kR)^{-3/2}]. \tag{4.14}$$

Since the first term in (4.14) is the dominant term in the asymptotic approximation of $u(x_0, y_0, z)$, we have from (1.1)

$$u_I(x_0, y_0, z) = -(i/kR)A(\pi/2) + O[(kR)^{-3/2}]. \tag{4.15}$$

Hence, $u_H(x_0, y_0, z)$ and $u_I(x_0, y_0, z)$ are of the same order in $(kR)^{-1}$ and in general $u_I(x_0, y_0, z)$ cannot be neglected compared to $u_H(x_0, y_0, z)$. In the special case, when $V(p, q, m)$ is of the form given in (1.13), however, $A(\pi/2) = 0$ and $u_I(x_0, y_0, z)$ can be neglected compared to $u_H(x_0, y_0, z)$ for large kR.

5. SUMMARY AND DISCUSSION

In this section, we give a summary of our main results to facilitate their application, and we discuss briefly their utility in some cases of interest when $U(p, q)$ does not belong to T_N.

A. Summary of results

Except where noted otherwise, all notation used here is defined in Sec. 1. The asymptotic approximations given are valid as $kR \to \infty$ for fixed ξ_1, ξ_2, k with $N \geq 12$. For less restrictive conditions on N in individual cases, see the text.

The asymptotic behavior of $u(x, y, z)$ is the same for all ξ_3 such that $0 \leq \xi_3 \leq 1$ [cf. (3.35)],

$$u(x, y, z) = -2\pi i[\exp(ikR)/kR][1 + (i/2kR)L^2] \times \exp[ik(\xi_1 x_0 + \xi_2 y_0 + \xi_3 z_0)]V(\xi_1, \xi_2, \xi_3) + O[(kR)^{-3}], \tag{5.1}$$

where L^2 is a differential operator defined by (2.54).

The asymptotic behavior of $u_H(x, y, z)$ depends on the value of ξ_3:

(a) $0 < \xi_3 < 1$ [cf. (4.5)]

$$u_H(x, y, z) = u(x, y, z) + u_{H1}(x, y, z) + O[(kR)^{-2}], \tag{5.2}$$

where $u_{H1}(x, y, z)$ is a term of order $(kR)^{-3/2}$ defined in (4.2),

(b) $\xi_3 = 0$ [cf. (4.7)],

$$u_H(x, y, z_0) = \tfrac{1}{2}u(x, y, z_0) + i\pi[\exp(-ikR)/kR] \times V(-\xi_1, -\xi_2, 0)\exp[-ik(x_0\xi_1 + y_0\xi_2)] + O[(kR)^{-3/2}], \tag{5.3}$$

(c) $\xi_3 = 1$ [cf. (4.14)],

$$u_H(x_0, y_0, z) = u(x_0, y_0, z) + (i/kR)\int_0^{2\pi} A(\alpha, \beta)\, d\beta + O[(kR)^{-3/2}], \tag{5.4}$$

where $A(\alpha, \beta)$ is given in (4.11).

The asymptotic behavior of $u_I(x, y, z)$ is obtained from (5.1)–(5.4) by applying $u_I(x, y, z) = u(x, y, z) - u_H(x, y, z)$. The results show that $u_I(x, y, z)$ is negligible compared to $u_H(x, y, z)$ for large kR in case (a), but not in cases (b) and (c).

In the special case when $V(p, q, m)$ is of the form given in (1.13), the first terms in (5.1) and (5.3) vanish. In that case, the dominant terms are given by (3.36) and (4.9).

B. Utility of results when $U(p,q)$ does not belong to T_N

In some applications of angular-spectrum representations, integrals of the form of $u(x, y, z)$ are obtained with $U(p, q)$ not belonging to T_N because of the presence of isolated singularities in the integrand. For example, see the analyses of the reflection and refraction of nonplanar waves at a plane interface in Gasper[20] and Stamnes.[21] The problem of obtaining asymptotic expansions of such integrals can be approached by using a neutralizer to isolate the singularity. The resulting integral that does not contain the singularity can be approximated by using the results of this paper since in that case $U(p, q) \in T_N$. This is one reason for allowing the functions $V(p, q, s)$ in the definition of T_N, to be nonanalytic because the presence of the neutralizer makes the function nonanalytic. (A neutralizer can not be an analytic function.)

In the remaining integral containing the singularity, $U(p, q)$ does not belong to T_N and must be dealt with in some other way. In some cases, however, such as in the reflection and refraction problems just cited, it is possible to change the variables of integration to transform the integral into one that can be approximated by the methods of this paper.

ACKNOWLEDGMENTS

We would like to express our gratitude to Professor Nicholas Chako and Professor Emil Wolf for helpful discussions concerning this problem.

APPENDIX A

As mentioned in Sec. 3, we have found formula (47) in Ref. 10, to be incorrect. We derive the correct formula in this appendix. The notation of Ref. 10 is used here without change except that we consistently make use of gamma functions rather than factorials, so that we use $\Gamma(z+1)$ where Jones and Kline[10] uses $z!$. Additional subscripts and superscripts are employed to indicate the integral and stationary point under consideration when we apply the result in Appendix B.

We first treat the case when $F_{2,0}$ and $F_{0,2}$ are positive. Then the expression for $h_0(t-F_{0,0})$ is given by (32) of Ref. 10, except that the limits of integration are from $-\pi/2$ to $\pi/2$,

$$h_0(t-F_{0,0})=\sum_{r=0}^{\infty}\frac{\partial^r}{\partial t^r}\int_{-\pi/2}^{\pi/2}\frac{1}{2(F_{2,0}F_{0,2})^{1/2}}\sum_{\mu=0}^{\infty}\sum_{\lambda=0}^{\mu}(t-F_{0,0})^{\mu/2}\times\frac{\cos^{\lambda}\eta\sin^{\mu-\lambda}\eta}{F_{2,0}^{\lambda/2}F_{0,2}^{(\mu-\lambda)/2}}\sum_{p=0}^{\lambda}\sum_{q=0}^{\mu-\lambda}G_{\lambda-p,\mu-\lambda-q}F_{r,p,q}\,d\eta\,. \tag{A1}$$

The integral and rth derivative appearing in (A1) can be expressed in terms of the gamma functions as follows

$$\int_{-\pi/2}^{\pi/2}\sin^{\mu-\lambda}\eta\cos^{\lambda}\eta\,d\eta=\tfrac{1}{2}[1+(-1)^{\mu-\lambda}]\frac{\Gamma[\frac{1}{2}(\mu-\lambda+1)]\Gamma[\frac{1}{2}(\lambda+1)]}{\Gamma(\frac{1}{2}\mu+1)}, \tag{A2}$$

$$\frac{\partial^r}{\partial t^r}(t-F_{0,0})^{\mu/2}=\begin{cases}\dfrac{\Gamma(\frac{1}{2}\mu+1)}{\Gamma(\frac{1}{2}\mu-r+1)}(t-F_{0,0})^{\mu/2-r}\\ \text{if } \mu \text{ is odd or } \mu \text{ is even and } \mu/2\geq r,\\ 0\\ \text{if } \mu \text{ is even and } r>\mu/2.\end{cases} \tag{A3}$$

Since the integral in (A2) vanishes if $\mu-\lambda$ is odd, we have two possibilities to consider, either μ and λ both even or μ and λ both odd. Let $h_0^{(1)}(t-F_{0,0})$ and $h_0^{(2)}(t-F_{0,0})$ be the contributions to $h_0(t-F_{0,0})$ due to even μ, λ and to odd μ, λ respectively. Similarly, let $J^{(1)}$ and $J^{(2)}$ be the contributions to the integral J due to $h_0^{(1)}(t-F_{0,0})$ and $h_0^{(2)}(t-F_{0,0})$ respectively.

We consider $h_0^{(1)}(t-F_{0,0})$ and $J^{(1)}$ first. Taking μ and λ to be even, substituting (A2) and (A3) into (A1), and changing summation variables so that μ, λ become 2μ, 2λ, we have

$$h_0^{(1)}(t-F_{0,0})=[1/2(F_{2,0}F_{0,2})^{1/2}]\sum_{r=0}^{\infty}\sum_{\mu=r}^{\infty}\frac{(t-F_{0,0})^{\mu-r}}{\Gamma(\mu-r+1)}\times\sum_{\lambda=0}^{\mu}\frac{\Gamma(\mu-\lambda+\frac{1}{2})\Gamma(\lambda+\frac{1}{2})}{F_{2,0}^{\lambda}F_{0,2}^{\mu-\lambda}}\times\sum_{p=0}^{2\lambda}\sum_{q=0}^{2(\mu-\lambda)}G_{2\lambda-p,2\mu-2\lambda-q}F_{r,p,q}. \tag{A4}$$

Now, let $\mu-r=m$. The coefficient of $[1/\Gamma(m+1)]\times(t-F_{0,0})^m$ is

$$\frac{1}{2(F_{2,0}F_{0,2})^{1/2}}\sum_{r=0}^{\infty}\sum_{\lambda=0}^{r+m}[\Gamma(\lambda+\tfrac{1}{2})\Gamma(m+r-\lambda+\tfrac{1}{2})/F_{2,0}^{\lambda}F_{0,2}^{m+r-\lambda}]\times\sum_{p=0}^{2\lambda}\sum_{q=0}^{2(m+r-\lambda)}G_{2\lambda-p,2m+2r-2\lambda-q}F_{r,p,q}.$$

Since $F_1^r(X,Y)=[F_{12}XY^2+F_{21}X^2Y+\cdots]^r$, it is clear that $F_{r,p,q}$ vanishes unless $p+q\geq 3r$. Hence, we have $3r\leq(p+q)_{\max}=2m+2r$, which implies that $F_{r,p,q}\neq 0$ only when $r\leq 2m$. The expression for $h_0^{(1)}(t-F_{0,0})$ becomes, therefore,

$$h_0^{(1)}(t-F_{0,0})=\frac{1}{2(F_{2,0}F_{0,2})^{1/2}}\sum_{m=0}^{\infty}\frac{(t-F_{0,0})^m}{\Gamma(m+1)}\times\sum_{r=0}^{2m}\sum_{\lambda=0}^{r+m}\frac{\Gamma(\lambda+\frac{1}{2})\Gamma(m+r-\lambda+\frac{1}{2})}{F_{2,0}^{\lambda}F_{0,2}^{m+r-\lambda}}S^{(1)}, \tag{A5}$$

$$S^{(1)}=\sum_{p=0}^{2\lambda}\sum_{q=0}^{2(m+r-\lambda)}G_{2\lambda-p,2m+2r-2\lambda-q}F_{r,p,q}. \tag{A6}$$

Applying Erdélyi's theorem (see Sec. 1 of Ref. 10), we have immediately

$$J^{(1)}\sim\frac{\exp(ikF_{0,0})}{2|F_{2,0}F_{0,2}|^{1/2}}\sum_{m=0}^{\infty}\frac{\exp[i(\pi/2)(m+1)]}{k^{m+1}}\times\sum_{r=0}^{2m}\sum_{\lambda=0}^{r+m}\frac{\Gamma(\lambda+\frac{1}{2})\Gamma(m+r-\lambda+\frac{1}{2})}{F_{2,0}^{\lambda}F_{0,2}^{m+r-\lambda}}S^{(1)}. \tag{A7}$$

Next we consider $h_0^{(2)}(t-F_{0,0})$ and $J^{(2)}$. Taking μ, λ to be odd, substituting (A2) and (A3) into (A1), and changing summation variables μ, λ to $2\mu+1$, $2\lambda+1$, we have

$$h_0^{(2)}(t-F_{0,0})=\frac{1}{2(F_{2,0}F_{0,2})^{1/2}}\sum_{r=0}^{\infty}\sum_{\mu=0}^{\infty}\frac{(t-F_{0,0})^{\mu-r+1/2}}{\Gamma(\mu-r+\frac{3}{2})}\times\sum_{\lambda=0}^{\mu}\frac{\Gamma(\mu-\lambda+\frac{1}{2})\Gamma(\lambda+1)}{F_{2,0}^{\lambda+1/2}F_{0,2}^{\mu-\lambda}}\times\sum_{p=0}^{2\lambda+1}\sum_{q=0}^{2(\mu-\lambda)}G_{2\lambda+1-p,2\mu-2\lambda-q}F_{r,p,q}. \tag{A8}$$

If we now let $\mu-r=m$ and note that in this case $F_{r,p,q}$ vanishes unless $3r\leq(p+q)_{\max}=2m+2r+1$, i.e., unless $0\leq r\leq 2m+1$, we may obtain

$$h_0^{(2)}(t-F_{0,0})=\frac{1}{2(F_{2,0}F_{0,2})^{1/2}}\sum_{m=0}^{\infty}\frac{(t-F_{0,0})^{m+1/2}}{\Gamma(m+\frac{3}{2})}\times\sum_{r=0}^{2m+1}\sum_{\lambda=0}^{r+m}\frac{\Gamma(m+r-\lambda+\frac{1}{2})\Gamma(\lambda+1)}{F_{2,0}^{\lambda+1/2}F_{0,2}^{m+r-\lambda}}S^{(2)}, \tag{A9}$$

where

$$S^{(2)}=\sum_{p=0}^{2\lambda+1}\sum_{q=0}^{2(m+r-\lambda)}G_{2\lambda+1-p,2m+2r-2\lambda-q}F_{r,p,q}. \tag{A10}$$

Application of Erdélyi's theorem then yields

$$J^{(2)}\sim\frac{\exp(ikF_{0,0})}{2|F_{2,0}F_{0,2}|^{1/2}}\sum_{m=0}^{\infty}\frac{\exp[i(\pi/2)(m+\frac{3}{2})]}{k^{m+3/2}}\times\sum_{r=0}^{2m+1}\sum_{\lambda=0}^{r+m}\frac{\Gamma(m+r-\lambda+\frac{1}{2})\Gamma(\lambda+1)}{F_{2,0}^{\lambda+1/2}F_{0,2}^{m+r-\lambda}}S^{(2)}. \tag{A11}$$

The total contribution to J due to a stationary point on the boundary is the sum $J^{(1)}+J^{(2)}$, with $J^{(1)}$ and $J^{(2)}$ as given by (A7) and (A11).

The expressions (A7) and (A11) are both derived under the condition that $F_{2,0}>0$ and $F_{0,2}>0$. If $F_{2,0}<0$ and $F_{0,2}<0$, then the sign of (A7) and (A11) should be reversed. If the stationary point is a saddlepoint such

that $f_{xx}f_{yy}-f_{xy}^2<0$ and hence, $F_{2,0}$ and $F_{0,2}$ have opposite signs, then $|F_{2,0}F_{0,2}|^{1/2}$ in (A7) and (A11) should be replaced by $i|F_{20}F_{02}|^{1/2}$. (In Sec. 6.2 of Ref. 10 it is said that $|F_{20}F_{02}|^{1/2}$ should be replaced by $|F_{20}F_{02}|^{1/2}/i$ in this case, but this is easily seen to give the wrong result for the first-order term.)

APPENDIX B

In this appendix, we determine the relationship between the coefficients in the expansion of $u_{H1}(x,y,z_0)$ and the coefficients in the expansion of $u_{I1}(x,y,z_0)$. As explained in the text, we can apply the results of Ref. 10 for that purpose even though the analysis of Ref. 10 is valid only for more restricted amplitude functions than those considered in the text. We simply treat the case when (in accordance with the restrictions imposed in Ref. 10) $V(p,q,s)$ is an analytic function of p, q, and s, knowing that the functional dependence of the coefficients is the same for all $U(p,q)$ in T_N. For the convenience of the reader, we use the notation of Ref. 10, i.e., the integration variables are denoted x,y and the amplitude and phase functions of the integrals are denoted $g(x,y)$ and $f(x,y)$ respectively. Additional subscripts and superscripts are used to indicate the integral and stationary point under consideration. The point of observation is denoted (x_{ob},y_{ob},z) in order to avoid confusion with the integration variables. We begin by considering u_{H1} as given by (3.12),

$$u_{H1}=\int_{\beta_0}^{\beta_0+2\pi}\int_{x_1'}^{\pi/2} g_H(x,y)\exp[ikRf_H(x,y)]\,dx\,dy, \tag{B1}$$

where

$$f_H(x,y)=\sin x\cos(y-\varphi),\quad \cos\varphi=\xi_1,\quad \sin\varphi=\xi_2, \tag{B2}$$

$$g_H(x,y)=V(p,q,m)\exp[ik(px_0+qy_0+mz_0)]\sin x, \tag{B3}$$

$$p=\sin x\cos y,\quad q=\sin x\sin y,\quad m=\cos x,$$
$$R=[(x_{ob}-x_0)^2+(y_{ob}-y_0)^2]^{1/2}. \tag{B4}$$

In writing (B3), we have omitted the neutralizer, it being understood that the boundary $x=x_1'$ gives no contributions. At the boundary $x=\pi/2$, where the critical points are, the neutralizer equals one.

The phase function $f_H(x,y)$ has two stationary points, point a at $x=\pi/2$, $y=\varphi$ and point b at $x=\pi/2$, $y=\varphi+\pi$. In what follows, we write $f^{(a)}_{Hm,n}$ and $g^{(a)}_{Hm,n}$ to represent the coefficients of the terms $(x-\pi/2)^m(y-\varphi)^n$ in the Taylor series expansions for f_H and g_H respectively centered at point a. Similarly, we write $f^{(b)}_{Hm,n}$ and $g^{(b)}_{Hm,n}$ for the coefficients in the corresponding expansions centered at point b. Application of the results of Jones and Kline[10] requires certain transformations in the integration variables. For the point j (where j is a or b), the variables of integration x,y are changed to $X_H^{(j)}$, $Y_H^{(j)}$ and the functions f_H, g_H are transformed to $F_H^{(j)}$, $G_H^{(j)}$. The coefficients of the terms $[X_H^{(j)}]^m[Y_H^{(j)}]^n$ in the Taylor series expansion for $F_H^{(j)}$ and $G_H^{(j)}$ centered at the origin are denoted $F^{(j)}_{Hm,n}$ and $G^{(j)}_{Hm,n}$ respectively. Similarly, the symbol $F^{(j)}_{Hr,m,n}$ is used to denote the coefficients of the terms $[X_H^{(j)}]^m[Y_H^{(j)}]^n$ in the Taylor series expansion of the quantity $[(-1)^r/r!][F_{H1}^{(j)}]^r$ where $F_{H1}^{(j)}=F_H^{(j)}-F^{(j)}_{H0,0}-F^{(j)}_{H2,0}[X_H^{(j)}]^2-F^{(j)}_{H0,2}[Y_H^{(j)}]^2$.

The coordinate transformations required in the present case are of a very simple nature. First of all, since the second-order cross-derivatives of the phase function vanish at the stationary points, no rotation of coordinate axes is needed. Secondly, since the boundary consists of straight lines parallel to the x or y axis, the second transformation of Ref. 10 (to make $\bar{\Phi}_{01}$ vanish) is unnecessary. The appropriate transformation for point a is therefore

$$x-\pi/2=-X_H^{(a)},\quad y-\varphi=-Y_H^{(a)}, \tag{B5}$$

and the transformation for point b is obtained by replacing φ by $\varphi+\pi$ in (B5). From (B5), it follows immediately that

$$F^{(j)}_{Hm,n}=(-1)^{m+n}f^{(j)}_{Hm,n},\quad G^{(j)}_{Hm,n}=(-1)^{m+n}g^{(j)}_{Hm,n}. \tag{B6}$$

Making use of the Taylor series expansions of $\sin x$ and $\cos x$ centered at point a and b, we find

$$f^{(j)}_{H2p+1,q}=f^{(j)}_{Hp,2q+1}=0,$$
$$f^{(a)}_{H2p,2q}=-f^{(b)}_{H2p,2q}=\frac{(-1)^{p+q}}{(2p)!\,(2q)!} \tag{B7}$$

It follows, then, from (B6) and (B7) that

$$F^{(j)}_{H2p+1,q}=F^{(j)}_{Hp,2q+1}=0,$$
$$F^{(a)}_{H2p,2q}=-F^{(b)}_{H2p,2q}=\frac{(-1)^{p+q}}{(2p)!\,(2q)!} \tag{B8}$$

and, in particular, that

$$F^{(a)}_{H0,0}=1,\quad F^{(a)}_{H2,0}=F^{(a)}_{H0,2}=-\tfrac{1}{2},$$
$$F^{(b)}_{H0,0}=-1,\quad F^{(b)}_{H2,0}=F^{(b)}_{H0,2}=\tfrac{1}{2}. \tag{B9}$$

We are now in a position to apply the results of Appendix A to obtain the desired coefficients. Substituting (B9) into (A7) and (A11) and taking into account the comment concerning sign in the last paragraph of Appendix A, we obtain the following expressions for the coefficients in the expansions in (3.17)—(3.18):

$$B_{Hm}^{(j1)}(\varphi)=\exp[i(\pi/2)(m+1)]\times\sum_{r=0}^{2m}\sum_{\lambda=0}^{r+m}\frac{\Gamma(\lambda+\frac{1}{2})\Gamma(m+r-\lambda+\frac{1}{2})}{(\frac{1}{2})^{m+r}}K_H^{(j1)}S_H^{(j1)}, \tag{B10}$$

$$B_{Hm}^{(j2)}(\varphi)=\exp[i(\pi/2)(m+3/2)]\times\sum_{r=0}^{2m+1}\sum_{\lambda=0}^{r+m}\frac{\Gamma(\lambda+1)\Gamma(m+r-\lambda+\frac{1}{2})}{(\frac{1}{2})^{m+r+1/2}}K_H^{(j2)}S_H^{(j2)}, \tag{B11}$$

where $S_H^{(j1)}$, $S_H^{(j2)}$ are defined by (A6), (A10), with G and F replaced by $G_H^{(j)}$ and $F_H^{(j)}$, and where $K_H^{(a1)}=(-1)^{m+r+1}$, $K_H^{(a2)}=(-1)^{m+r+1/2}$, and $K_H^{(b1)}=K_H^{(b2)}=1$.

We now proceed to consider u_{I1} as given by (3.23),

$$u_{I1}=\int_{\beta_0}^{\beta_0+2\pi}\int_0^{x_1} g_I(x,y)\exp[ikRf_I(x,y)]\,dx\,dy, \tag{B12}$$

where

$$f_I(x,y)=\cosh x\cos(y-\varphi), \tag{B13}$$

$$g_I(x,y)=-iV(p,q,m)\exp[ik(px_0+qy_0+mz_0)]\cosh x, \tag{B14}$$

$$p=\cosh x\cos y,\quad q=\cosh x\sin y,$$
$$m=i\sinh x. \tag{B15}$$

The neutralizer has again been dropped for the same reasons as before. The two stationary points a and b are located at $x=0$, $y=\varphi$ and $x=0$, $y=\varphi+\pi$. In what follows, we use the same notation as above with H replaced by I.

The coordinate transformations for a are given by

$$x=X_I^{(a)}, \quad y-\varphi=Y_I^{(a)}, \tag{B16}$$

and the transformation for point b is obtained by replacing φ by $\varphi+\pi$ in (B16). We immediately have

$$F_{Im,n}^{(j)}=f_{Im,n}^{(j)}, \quad G_{Im,n}^{(j)}=g_{Im,n}^{(j)}. \tag{B17}$$

Making use of the Taylor series expansions of $\cosh x$, $\sin y$, $\cos y$, we find the expansion coefficients

$$f_{I2p+1,q}^{(j)}=f_{Ip,2q+1}^{(j)}=0, \quad f_{I2p,2q}^{(j)}=(-1)^p f_{H2p,2q}^{(j)}. \tag{B18}$$

It follows from (B17) and (B18), that

$$F_{I2p+1,q}^{(j)}=F_{Ip,2q+1}^{(j)}=0, \quad F_{I2p,2q}^{(j)}=f_{I2p,2q}^{(j)}=(-1)^p f_{H2p,2q}^{(j)}, \tag{B19}$$

and in particular that

$$\begin{aligned} &F_{I0,0}^{(a)}=1, \quad F_{I2,0}^{(a)}=-F_{I0,2}^{(a)}=\tfrac{1}{2}, \\ &F_{I0,0}^{(b)}=-1, \quad F_{I2,0}^{(b)}=-F_{I0,2}^{(b)}=-\tfrac{1}{2}. \end{aligned} \tag{B20}$$

Substituting (B20) into (A7) and (A11) and taking into account the fact that the points a and b are saddlepoints, we find that the expressions for the coefficients $B_{Im}^{(jl)}(\varphi)$ (with $j=a, b$ and $l=1, 2$) are identical to those for $B_{Hm}^{(jl)}(\varphi)$ in (B10)—(B11), except that $K_H^{(jl)}$ is replaced by $K_I^{(jl)}$ where $K_I^{(a1)}=i(-1)^\lambda K_H^{(a1)}$, $K_I^{(a2)}=-(-1)^\lambda K_H^{(a2)}$, $K_I^{(b1)}=-i(-1)^\lambda K_H^{(b1)}$, and $K_I^{(b2)}=-(-1)^\lambda K_H^{(b2)}$.

To obtain the asymptotic behavior of $u(x, y, z_0)$, we must add the asymptotic expansions of $u_{H1}^{(a1)}$, $u_{H1}^{(a2)}$, $u_{H1}^{(b1)}$, $u_{H1}^{(b2)}$, $u_{I1}^{(a1)}$, $u_{I1}^{(a2)}$, $u_{I1}^{(b1)}$ and $u_{I1}^{(b2)}$. Since the terms in the asymptotic expansions of $u_{H1}^{(jl)}$ (with $j=a, b$ and $l=1, 2$) are identical to those of $u_I^{(jl)}$ apart from the coefficients $B_{Hm}^{(jl)}$ and $B_{Im}^{(jl)}$, the asymptotic approximation of $u(x, y, z_0)$ becomes the sum of four series, each of which has coefficients of the form $B_{Hm}^{(jl)}+B_{Im}^{(jl)}$. Hence, we are interested in the following quantities:

$$\begin{aligned} B_{Hm}^{(a1)}+B_{Im}^{(a1)} &= \exp[i(\pi/2)(m+1)] \\ &\times \sum_{r=0}^{2m}\sum_{\lambda=0}^{r+m}\frac{\Gamma(\lambda+\frac{1}{2})\Gamma(m+r-\lambda+\frac{1}{2})}{(\frac{1}{2})^{m+r}}K_H^{(a1)} \\ &\times[S_H^{(a1)}+i(-1)^\lambda S_I^{(a1)}], \end{aligned} \tag{B21}$$

$$\begin{aligned} B_{Hm}^{(b1)}+B_{Im}^{(b1)} &= \exp[i(\pi/2)(m+1)]\sum_{r=0}^{2m}\sum_{\lambda=0}^{r+m}\frac{\Gamma(\lambda+\frac{1}{2})\Gamma(m+r-\lambda+\frac{1}{2})}{(\frac{1}{2})^{m+r}} \\ &\times[S_H^{(b1)}-i(-1)^\lambda S_I^{(b1)}], \end{aligned} \tag{B22}$$

$$\begin{aligned} B_{Hm}^{(j2)}+B_{Im}^{(j2)} &= \exp[i(\pi/2)(m+\tfrac{3}{2})]\sum_{r=0}^{2m+1}\sum_{\lambda=0}^{r+m}\frac{\Gamma(\lambda+1)\Gamma(m+r-\lambda+\frac{1}{2})}{(\frac{1}{2})^{m+r+1/2}}K_H^{(j2)} \\ &\times[S_H^{(j2)}-(-1)^\lambda S_I^{(j2)}]. \end{aligned} \tag{B23}$$

The quantities in the final square brackets [ ] in (B21)—(B23) can be simplified by comparing the products $G_{H2\lambda-p,2m+2r-2\lambda-q}^{(j)}F_{Hr,p,q}^{(j)}$ and $G_{H2\lambda+1-p,2m+2r-2\lambda-q}^{(j)}F_{Hr,p,q}^{(j)}$ to the corresponding products with H replaced by I.

We begin by considering the terms $F_{Jr,p,q}^{(j)}$ with $j=a, b$ and $J=H, I$. We have

$$\begin{aligned} \sum_{p,q=0}^{\infty}F_{Jr,p,q}^{(j)}[X_J^{(j)}]^p[Y_J^{(j)}]^q &= \frac{(-1)^r}{r!}[F_{J1}^{(j)}]^r \\ &= \frac{(-1)^r}{r!}\left(\sum_{m,n=0}^{\infty}F_{Jm,n}^{(j)}[X_J^{(j)}]^m[Y_J^{(j)}]^n\right)^r. \end{aligned} \tag{B24}$$

Hence, $[r!/(-1)^r]F_{Jr,p,q}^{(j)}$ is a sum of terms, each of which is a product of r factors, $F_{Jm,n}^{(j)}$, with various values of m, n. In any one term, the sum of the indices labeled m must be p, and the sum of the indices labeled n must be q. Hence, it follows from (B6) that each term in $[r!/(-1)^r]F_{Hr,p,q}^{(j)}$ is the product of $(-1)^{p+q}$ with r factors, $f_{Hm,n}^{(j)}$. Similarly it follows from (B17) and (B19) that if we replace H by I in the expression for $[r!/(-1)^r]\times F_{Hr,p,q}^{(j)}$, each term is unchanged except that the factor $(-1)^{p+q}$ becomes $(-1)^{p/2}$. Hence we have

$$F_{Hr,p,q}^{(j)}=(-1)^{(p/2)+q}F_{Ir,p,q}^{(j)}. \tag{B25}$$

Also, we note that if p or q is odd, then each term in the expression for $[r!/(-1)^r]F_{Hr,p,q}^{(j)}$ contains at least one factor $F_{Hm,n}^{(j)}$ with odd m or n. Since according to (B7), any such factor is zero, it follows that

$$F_{Jr,p,2q+1}^{(j)}=F_{Jr,2p+1,q}^{(j)}=0. \tag{B26}$$

Hence, we need consider only even p, q. (B25) becomes

$$F_{Hr,2p,2q}^{(j)}=(-1)^p F_{Ir,2p,2q}^{(j)}. \tag{B27}$$

Next we compare the coefficients $G_{Hm,n}^{(j)}$ and $G_{Im,n}^{(j)}$. From (B6) and (B17), we have

$$\frac{G_{Hm,n}^{(j)}}{G_{Im,n}^{(j)}}=(-1)^{m+n}\frac{g_{Hm,n}^{(j)}}{g_{Im,n}^{(j)}}. \tag{B28}$$

The quantity $g_{Jm,n}^{(j)}$ (with $J=H, I$) is proportional (with proportionality constant independent of J and j) to $\partial^{m+n}g_J/\partial x^m\partial y^n$ evaluated at point j. Consequently, the ratios appearing in (B28) can be obtained by finding the ratios of the corresponding partial derivatives. To that end, we rewrite (B3) and (B14) as

$$g_H(x, y)=B(S_{Hx}, C_{Hx}, S_y, C_y), \tag{B29}$$

$$g_I(x, y)=(1/i)B(S_{Ix}, C_{Ix}, S_y, C_y), \tag{B30}$$

where

$$S_{Hx}=\sin x, \quad C_{Hx}=\cos x, \quad S_{Ix}=\cosh x, \tag{B31}$$

$$C_{Ix}=i\sinh x, \quad S_y=\sin y, \quad C_y=\cos y, \tag{B32}$$

and (with $J=H, I$)

$$\begin{aligned} B(S_{Jx}, C_{Jx}, S_y, C_y) &= V(p, q, m)\exp[ik(px_0+qy_0+mz_0)] \\ &\times S_{Jx}, \end{aligned} \tag{B33}$$

with

$$p=S_{Jx}C_y, \quad q=S_{Jx}C_y, \quad m=C_{Jx}. \tag{B34}$$

In terms of the new variables, the points a and b are the points $S_{Jx}=1$, $C_{Jx}=0$, $S_y=\sin\varphi$, $C_y=\cos\varphi$ and $S_{Jx}=1$, $C_{Jx}=0$, $S_y=\sin(\varphi+\pi)$, $C_y=\cos(\varphi+\pi)$, respectively.

Now let Q be the set of all pairs of functions A_H, A_I of the form

$$A_H = A(S_{Hx}, C_{Hx}, S_y, C_y), \tag{B35}$$

$$A_I = A(S_{Ix}, C_{Ix}, S_y, C_y), \tag{B36}$$

where $A(S_{Jx}, C_{Jx}, S_yC_y)$ has continuous partial derivatives of all orders (with respect to S_{Jx}, C_{Jx}, S_y, C_y treated as independent variables) at points a and b. Then, we have

$$\frac{\partial A_H}{\partial x} = C_{Hx}\frac{\partial A_H}{\partial S_{Hx}} - S_{Hx}\frac{\partial A_H}{\partial C_{Hx}}, \tag{B37}$$

$$\frac{\partial A_I}{\partial x} = \frac{1}{i}\left(C_{Ix}\frac{\partial A_I}{\partial S_{Ix}} - S_{Ix}\frac{\partial A_I}{\partial x}\right). \tag{B38}$$

Hence, the pair $\partial A_H/\partial x$, $i\partial A_I/\partial x$ belongs to Q. Repetition of the above procedure m times shows that the pair $\partial^m A_H/\partial x^m$, $i^m\partial^m A_I/\partial x^m$ also belongs to Q. Next, by applying the same procedure with the x differentiation replaced by y differentiation we find that the pair $\partial^{m+n}A_H/\partial x^m\partial y^n$, $i^m\partial^{m+n}A_I/\partial x^m\partial y^n$ belongs to Q. Finally, we note that if a pair of functions belongs to Q, then the functions in the pair are equal to each other at point a and point b. Hence, we have

$$\frac{\partial^{m+n}A_H}{\partial x^m\partial y^n} = i^m\frac{\partial^{m+n}A_I}{\partial x^m\partial y^n} \tag{B39}$$

at points a and b. Since, according to (B29)—(B30), the pair $g_H(x, y)$, $ig_I(x, y)$ belongs to Q, it follows from (B39) that

$$\frac{\partial^{m+n}}{\partial x^m\partial y^n}g_H(x, y) = i^{m+1}\frac{\partial^{m+n}}{\partial x^m\partial y^n}g_I(x, y) \tag{B40}$$

at points a and b. Hence (B28) becomes

$$G^{(j)}_{Hm,n} = (-1)^{m+n}i^{m+1}G^{(j)}_{Im,n}. \tag{B41}$$

We are now ready to examine the terms $S_J^{(j1)}$ and $S_J^{(j2)}$. By employing (B26) in (A6) and (A10), we have

$$S_J^{(j1)} = \sum_{p=0}^{\lambda}\sum_{q=0}^{m+r-\lambda} G^{(j)}_{J\,2(\lambda-p),2(m+r-\lambda-q)}F^{(j)}_{Jr,2p,2q}, \tag{B42}$$

$$S_J^{(j2)} = \sum_{p=0}^{\lambda}\sum_{q=0}^{m+r-\lambda} G^{(j)}_{J\,2(\lambda-p)+1,2(m+r-\lambda-q)}F^{(j)}_{Jr,2p,2q}. \tag{B43}$$

Combination of (B27) and (B41) yields

$$F^{(j)}_{Hr,2p,2q}G^{(j)}_{Hm,n} = i^{m+1}(-1)^{m+n+p}F^{(j)}_{Ir,2p,2q}G^{(j)}_{Im,n}. \tag{B44}$$

When we apply (B44) in (B42) and (B43), we find

$$S_H^{(j1)} = i(-1)^{\lambda}S_I^{(j1)}, \tag{B45}$$

$$S_H^{(j2)} = (-1)^{\lambda}S_I^{(j2)} \tag{B46}$$

Finally, we substitute (B45)—(B46) into (B21)—(B23) to obtain

$$B^{(a1)}_{Hm} + B^{(a1)}_{Im} = 2B^{(a1)}_{Hm}, \tag{B47}$$

$$B^{(b1)}_{Hm} + B^{(b1)}_{Im} = 0, \tag{B48}$$

$$B^{(j2)}_{Hm} + B^{(j2)}_{Im} = 0. \tag{B49}$$

These equations are equivalent to (3.28)—(3.31), the results to be obtained in this appendix.

APPENDIX C

In this appendix we derive (3.34) of the text, i.e., we show that

$$\lim_{\theta\to\pi/2} B_n(\theta, \varphi) = 2B^{(a2)}_{Hn}(\varphi). \tag{C1}$$

The notation of the previous appendix, which is close to that of Jones and Kline[10] is employed.

Using (3.17), (A6), and (A7) (with k replaced by kR) and taking account of the comment concerning sign in the last paragraph of Appendix A, we find that the coefficients $B^{(a1)}_{Hn}(\varphi)$ are given by

$$B^{(a1)}_{Hn}(\varphi) = -\frac{\exp[ikR(F^{(a)}_{H0,0}-1)]\exp[i\pi(n+1)/2]}{2|F^{(a)}_{H2,0}F^{(a)}_{H0,2}|^{1/2}} \times\sum_{r=0}^{2n}\sum_{\lambda=0}^{r+n}\frac{\Gamma(\lambda+\frac{1}{2})\Gamma(n+r-\lambda+\frac{1}{2})}{[F^{(a)}_{H2,0}]^{\lambda}[F^{(a)}_{H0,2}]^{n+r-\lambda}} \times\sum_{p=0}^{2\lambda}\sum_{q}^{2(n+r-\lambda)} G^{(a)}_{H2\lambda-p,2n+2r-2\lambda-q}F^{(a)}_{Hr,p,q}, \tag{C2}$$

where the phase and amplitude functions have the form

$$f_H(x, y) = \sin x\cos(y-\varphi) \tag{C3}$$

and

$$g_H(x, y) = V(\sin x\cos y,\ \sin x\sin y,\ \cos x) \times\exp[ik(x_0\sin x\cos y + y_0\sin x\sin y + z_0\cos x)] \times\sin x, \tag{C4}$$

respectively (the constants $F^{(a)}_{Hi,j}$, $F^{(a)}_{Hi,m,n}$ and $G^{(a)}_{H,i,j}$ are derived from the phase and amplitude functions by the procedure indicated in Ref. 10) and where R is defined by

$$R = [(x_{ob}-x_0)^2 + (y_{ob}-y_0)^2]^{1/2}, \tag{C5}$$

the point (x_{ob}, y_{ob}, z) being the point of observation.

The coefficient $B_n(\theta, \varphi)$, on the other hand, may be found using (2.51) of the text and (3.4) of Ref. 10 (with the factorials replaced by Γ functions, k replaced by kR' and the overall sign reversed since $F_{2,0} < 0$ and $F_{0,2} > 0$ in the present instance); they are given by

$$B_n(\theta, \varphi) = -\frac{\exp[ikR'(F_{0,0}-1)]\exp[i(\pi/2)(n+1)]}{|F_{2,0}F_{0,2}|^{1/2}} \times\sum_{r=0}^{2n}\sum_{\lambda=0}^{r+n}\frac{\Gamma(\lambda+\frac{1}{2})\Gamma(n+r-\lambda+\frac{1}{2})}{F^{\lambda}_{2,0}F^{n+r-\lambda}_{0,2}} \times\sum_{p=0}^{2\lambda}\sum_{q=0}^{2(n+r-\lambda)} G_{2\lambda-p,2n+2r-2\lambda-q}F_{r,p,q}, \tag{C6}$$

where in this case, the phase and amplitude functions may be expressed in the form

$$f(x, y) = \sin\theta\sin x\cos(y-\varphi) + \cos\theta\sin y \tag{C7}$$

and

$$g(x, y) = V(\sin x\cos y,\ \sin x\sin y,\ \cos x) \times\exp[ik(x_0\sin x\cos y + y_0\sin x\sin y + z_0\cos x)\sin x], \tag{C8}$$

respectively, and

$$R' = [(x_{ob}-x_0)^2 + (y_{ob}-y_0)^2 + (z-z_0)^2]^{1/2}. \tag{C9}$$

In writing (C4) and (C8), we have omitted the neu-

tralizers for the same reason as given in Appendix B. We have also taken into account, in writing (C2) and (C6), that the stationary point under consideration is, in both cases, a relative maximum.

A comparison of (C4) and (C8) shows that

$$g(x, y) = g_H(x, y), \tag{C10}$$

while a comparison of (C3) and (C7) yields

$$\lim_{\theta \to \pi/2} f(x, y) = f_H(x, y). \tag{C11}$$

From (C4) and (C8) and the definition of $g_{m,n}$ and $g^{(a)}_{Hm,n}$ we may obtain

$$\begin{aligned}\lim_{\theta \to \pi/2} & g_{m,n}(\theta, \varphi) \\ &\equiv \frac{1}{m!\,n!} \frac{\partial^{m+n} g(x, y)}{\partial x^m \partial y^n}\bigg|_{x=\theta,\, y=\varphi} \\ &= \frac{1}{m!\,n!} \frac{\partial^{m+n} g_H(x, y)}{\partial x^m \partial y^n}\bigg|_{x=\pi/2,\, y=\varphi} \equiv g^{(a)}_{Hm,n}(\varphi);\end{aligned} \tag{C12}$$

similarly using (C3), (C7), and the definition of $f_{m,n}$ and $f^{(a)}_{Hm,n}$ we find that

$$\begin{aligned}\lim_{\theta \to \pi/2} & f_{m,n}(\theta, \varphi) \\ &= \frac{1}{m!\,n!} \frac{\partial^{m+n} f(x, y)}{\partial x^m \partial y^n}\bigg|_{x=\theta,\, y=\varphi} \\ &= \frac{1}{m!\,n!} \frac{\partial^{m+n} f_H(x, y)}{\partial x^m \partial y^n}\bigg|_{x=\pi/2,\, y=\varphi} \equiv f^{(a)}_{Hm,n}(\varphi).\end{aligned} \tag{C13}$$

It should be noted that $g_{m,n}$ is a function of θ since the derivatives are evaluated at the stationary point ($x = \theta$, $y = \varphi$) whereas $f_{m,n}$ is a function of θ both because of the explicit dependence of $f(x, y)$ on θ and because the derivatives are evaluated at the stationary point.

It may also be seen immediately that since $z \to z_0$ as $\theta \to \pi/2$, we have

$$\lim_{\theta \to \pi/2} R' = R. \tag{C14}$$

It follows readily from (C7) that the second-order cross-derivative of $f(x, y)$ vanishes at the stationary point, i.e., $\partial^2 f/\partial x \partial y\,|_{x=\theta,\, y=\varphi} = 0$. The phase function $f(x, y)$ is therefore already in the form given by (30) of Ref. 10 and so we have

$$f_{m,n}(\theta, \varphi) = F_{m,n}(\theta, \varphi), \quad g_{m,n}(\theta, \varphi) = G_{m,n}(\theta, \varphi). \tag{C15}$$

On the other hand we find directly from (B6) that

$$f^{(a)}_{Hm,n}(\varphi) = (-1)^{m+n} F^{(a)}_{Hm,n}(\varphi), \quad g^{(a)}_{Hm,n}(\varphi) = (-1)^{m+n} G^{(a)}_{Hm,n}(\varphi). \tag{C16}$$

If we now use (C12), (C13), (C15) and (C16), we obtain

$$\lim_{\theta \to \pi/2} F_{m,n}(\theta, \varphi) = (-1)^{m+n} F^{(a)}_{Hm,n}(\varphi) \tag{C17}$$

and

$$\lim_{\theta \to \pi/2} G_{m,n}(\theta, \varphi) = (-1)^{m+n} G^{(a)}_{Hm,n}(\varphi). \tag{C18}$$

We know from (B8) that $F^{(a)}_{Hm,n}(\varphi)$ vanishes unless m and n are even; therefore, by (C17) the same statement holds for $\lim_{\theta \to \pi/2} F_{m,n}(\theta, \varphi)$, and we have

$$\lim_{\theta \to \pi/2} F_{2m,2n}(\theta, \varphi) = F^{(a)}_{H2m,2n}(\varphi). \tag{C19}$$

Because of the definition of $F_{r,m,n}$ and $F^{(a)}_{Hr,m,n}$, it follows from the last equation that

$$\lim_{\theta \to \pi/2} F_{r,2p,2q}(\theta, \varphi) = F^{(a)}_{Hr,2p,2q}(\varphi), \tag{C20}$$

the terms with odd subscripts in either of the last two places being zero. We may now see from the form of (C2) and (C6) that $G_{m,n}$ and $G^{(a)}_{Hm,n}$ contribute to these expressions only when m and n are even. In this case, we have from (C18)

$$\lim_{\theta \to \pi/2} G_{2m,2n}(\theta, \varphi) = G^{(a)}_{H2m,2n}(\varphi). \tag{C21}$$

Finally, using (C14), (C19), (C20), and (C21) in (C2) and (C6), we obtain

$$\lim_{\theta \to \pi/2} B_n(\theta, \varphi) = 2B^{(a1)}_{Hn}(\varphi), \tag{C22}$$

which was to be proved.

*Research supported by the U.S. Air Force Office of Scientific Research and the Army Research Office (Durham). A summary of the main result obtained herein is presented in G.C. Sherman, J.J. Stamnes, A.J. Devaney, and É. Lalor, Opt. Commun. 8, 271 (1973).

†Present affiliation: Norwegian Defense Research Establishment, Division for Electronics, P.O. Box 25, 2007 Kjeller, Norway.

‡Present affiliation: National Science Council, St. Martin's House, Waterloo Road, Dublin 4, Ireland.

[1]C.J. Bouwkamp, Rep. Prog. Phys. 17, 39 (1954).
[2]É. Lalor, J. Opt. Soc. Am. 58, 1235 (1968).
[3]A.J. Devaney and G.C. Sherman, SIAM Rev. 15, 765 (1973).
[4]K. Miyamoto and E. Wolf, J. Opt. Soc. Am. 52, 615 (1962).
[5]H.M. Nussenzveig, An. Acad. Bras. Ciênc. 31, 515 (1959).
[6]J.G. van der Corput, Compos. Math. 1, 15 (1934).
[7]A. Erdélyi, J. Soc. Ind. Appl. Math. 3, 17 (1955).
[8]J. Focke, Ber. Verh. Saechs. Akad. Wiss. Leipzig, 101, 1 (1954).
[9]G. Braun, Acta Phys. Austriaca 10, 8 (1956).
[10]D.S. Jones and M. Kline, J. Math. Phys. 37, 1 (1958).
[11]N. Chako, J. Inst. Math. Its Appl. 1, 372 (1965).
[12]A. Baños, *Dipole Radiation in the Presence of a Conducting Half-Space* (Pergamon, N.Y., 1966).
[13]P.C. Clemmow, *Plane Wave Spectrum Representation of Electromagnetic Fields* (Pergamon, N.Y., 1966).
[14]G. Tyras, *Radiation and Propagation of Electromagnetic Waves* (Academic, N.Y., 1969).
[15]A. Baños, *Selected Topics on Asymptotic Methods*, lecture notes available from Dept. of Physics, University of California at Los Angeles, 1969.
[16]P.M. Morse and H. Feshbach, *Methods of Theoretical Physics* (McGraw-Hill, N.Y., 1953), pp. 1539—40.
[17]Hans Bremermann, *Distributions, Complex Variables and Fourier Transforms* (Addison-Wesley, Reading, Mass., 1965), Sec. 3.3—3.4.
[18]J.J. Stamnes and G.C. Sherman, to be published.
[19]G.C. Sherman, Radio Sci. 8, 811 (1973).
[20]J. Gasper, M.S. Thesis, University of Rochester, 1972.
[21]J.J. Stamnes, Ph.D. Thesis, University of Rochester, 1974.

Geometry of hyperspace. I*†

Karel Kuchař

Department of Physics, University of Utah, Salt Lake City, Utah 84112
(Received 6 June 1975)

Hyperspace is heuristically defined as an (infinitely dimensional) manifold of all spacelike hypersurfaces embedded in a given Riemannian spacetime. The Riemannian structure $(\mathcal{M},\underline{g})$ of spacetime induces a rich geometrical structure in hyperspace. Part of that structure, especially the moving normal frames in hyperspace, Lie derivatives, and symmetrical ∇ and asymmetrical $\overset{*}{\nabla}$ covariant hyperderivatives, are studied in detail. The formalism introduced in this paper sets the stage for the geometrical study of hypersurface kinematics and dynamics of general tensor fields with derivative gravitational coupling, and of the Dirac–ADM geometrodynamics with such tensor sources, in the following papers.

1. INTRODUCTION

The dynamical evolution of a tensor field in a Riemannian spacetime is conveniently studied by cutting the spacetime manifold into a foliation of spacelike hypersurfaces, projecting the tensor field into tangential and normal directions to the hypersurfaces, and following how these projections change from one hypersurface to another. The same method can be used for studying the evolution of the geometry itself. The evolution is then concisely described in terms of the generalized canonical formalism. Such an approach to field dynamics was initiated by Dirac[1] and by Arnowitt, Deser, and Misner (ADM),[2] who applied it to the gravitational field and to simple tensor fields with nonderivative gravitational coupling (the scalar and the electromagnetic fields). The general theory of tensor fields with a derivative gravitational coupling was never fully developed along these lines.

Both Dirac and ADM use a coordinate-dependent language,[3] in which the hypersurfaces in the foliation are characterized as the coordinate hypersurfaces $T=X^0=\text{const}$ in a spacetime coordinate system X^α, and the coordinates X^a are used to label the points of the hypersurfaces. In such a language, the change of spacetime coordinates $X^\alpha \to X^{\alpha'}(X^\beta)$ induces the change of the foliation. The geometrical meaning of the Dirac–ADM field dynamics, which is essentially derived from the theory of embeddings, gets thus largely obscured.

Even when one firmly sticks to the language of foliations and embeddings, the situation is not fully satisfactory. The foliation is picked out either by "coordinate conditions" (ADM), or by hand (Dirac). The full arbitrariness in the choice of foliation never finds its proper expression. In this connection, statements like "canonical formalism necessarily destroys the spacetime covariance of the theory" are often heard. The terminology contrasting "the canonical theory" with "the covariant theory"[4] seems to point in the same direction, namely, that the canonical theory is not (at least "manifestly") covariant. Such feelings, however, are founded on a particular and rather unfortunate choice of formal language, rather than on the nature of the subject.

In this paper, we propose a formalism which tries to avoid these shortcomings.[5] Its essence may be summarized by saying that the field dynamics is not properly described as taking place in spacetime, or along a single foliation of hypersurfaces drawn in spacetime, but in hyperspace. Heuristically, hyperspace is the (infinitely dimensional) manifold of all spacelike hypersurfaces drawn in a given Riemannian spacetime. In it, all spacelike hypersurfaces are democratically put on the same footing. Foliation of spacelike hypersurfaces is a curve in hyperspace (though not every curve in hyperspace is a foliation, because the intersection of individual hypersurfaces is allowed for curves, but not for foliations). Infinitesimal deformation of a hypersurface is a tangent vector to a curve in hyperspace. Projections of a tensor field parallel and perpendicular to a hypersurface form a fiber over this hypersurface. *Hypersurface dynamics* studies how the field point in this fiber changes under the displacement of the base point in hyperspace.

Hyperspace has a rich geometrical structure which is largely inherited from spacetime. In this paper, we explore that part of the structure which is connected with the concepts of hyperspace, the tangent space to hyperspace, the covariant differentiation in hyperspace, and its extension to the "bundle of e-tensors." In subsequent papers, we build a general dynamical theory of tensor fields propagating in a given Riemannian spacetime or coupled to that spacetime by Einstein's law of gravitation, as a dynamical theory of field projections in hyperspace.

We start by summarizing our notation for any future reference (Sec. 2), as many analogous operations (like the Lie or the covariant derivatives) in different spaces which are introduced in our study (space, spacetime, the space of embeddings, hyperspace, the bundle of e-tensors) need to be distinguished from each other. In Sec. 3, we review the basic material from the theory of embeddings (vectors tangent and normal to the hypersurface, induced metric and affine connection, projections of tensors $\parallel$ and $\perp$ to the hypersurface, extrinsic curvature). In Sec. 4, we heutristically introduce the concepts of the space of embeddings $\mathcal{E}$ as an infinitely dimensional manifold $\mathcal{E}$ of all embeddings e which lead to a spacelike hypersurface in spacetime, and of hyperspace as the quotient space of $\mathcal{E}$ by the group of space diffeomorphisms. We study the character of the tangent space $T_e(\mathcal{E})$ to the space of embeddings, and define the "normal $\mathcal{E}$-basis" in $T_e(\mathcal{E})$, which leads to the standard lapse-shift decomposition of the deformation $\mathcal{E}$-vector. In Sec. 5, we introduce the bundle of e-tensors over $\mathcal{E}$. An e-tensor is essentially a field of

mixed spacetime–space tensors defined along an embedding and considered as a single object. Spacetime tensor fields intersected by an embedding and projected into tangential and normal directions to the hypersurface are special examples of e-tensors. Lie brackets between two $\mathcal{E}$-vectors are studied in Sec. 6. In particular, the Lie brackets of any two $\mathcal{E}$-vectors, δ_{Lx} and δ_{ax}, from the normal $\mathcal{E}$-basis, are evaluated, and it is shown that they close in the same way as the super-Hamiltonian and supermomentum of a dynamical theory (compare with Ref. 6). In Sec. 7, we discuss another Lie-type operation, the derivative $\mathbf{L}_{\tilde{N}}$ of an e-tensor field along a space vector field. If this derivative vanishes, the e-tensor may be interpreted as a hypertensor, being the same for all embeddings which lead to the same hypersurface. Similarly, we get a criterion for an $\mathcal{E}$-vector to be an $\mathcal{H}$-vector, i.e., a tangent vector in hyperspace. Covariant differentiation of spacetime vectors directly induces a symmetrical covariant differentiation ∇ in the space of embeddings, which can be extended from $\mathcal{E}$-vector fields to e-tensor fields. Moreover, due to its commutation with $\mathbf{L}_{\tilde{N}}$, this covariant derivative may be interpreted as a covariant derivative in hyperspace. The covariant derivative ∇ of the tangential and normal hyperfields is evaluated in Sec. 9. It characterizes the deformation of the normal hyperbasis under the deformation of the hypersurface. The normal deformations of the hyperbasis may be used to classify the deformations locally into "tilts" and "translations" (Sec. 10). This is important later, in the hypersurface kinematics of tensor fields, as the behavior of an e-field under the tilt shows whether the e-field can be interpreted as an intersection plus a projection of a spacetime tensor field. Hypersurface tilts lead to the terms in the field super-Hamiltonian which are nonlocal in the field momenta, and are closely connected with the spin energy-momentum tensor of the field. The induced covariant derivative ∇ does not commute with the projections and with the raising and lowering of the space tensor indices. In Sec. 11, a new (nonsymmetrical) covariant derivative $\overset{*}{\nabla}$ is discussed, which has these desirable properties. This covariant derivative preserves the normal hyperbasis vectors and the space and spacetime metric tensors, considered as hypertensor fields. Either one of the covariant derivatives, ∇ or $\overset{*}{\nabla}$, helps to transfer the spacetime covariant derivatives of spacetime tensor fields into hyperspace derivatives of the field projections, which is the basic trick used when rewriting the spacetime field dynamics into the field dynamics in hyperspace.

The formalism introduced in this paper is applied to the geometrical formulation of hypersurface kinematics and dynamics of general tensor fields with a derivative gravitational coupling, and of geometrodynamics with such tensor sources. This is the subject of the following papers, "Kinematics of Tensor Fields in Hyperspace (II)," "Dynamics of Tensor Fields in Hyperspace (III)," and "Tensor Sources in Geometrodynamics (IV)."

2. NOTATION

Manifold	Point	Local coordinates, components, basis
m... space manifold	$x \in m$	x^a, $a = 1, 2, 3$
$\mathcal{M}$... spacetime manifold	$X \in \mathcal{M}$	X^α, $\alpha = 0, 1, 2, 3$
$\mathcal{E}$... space of embeddings [Eq. (4.1)]	$e \in \mathcal{E}$	$e^\alpha(x^a)$
$\mathcal{H}$... hyperspace [Eq. (4.3)]	$h \in \mathcal{H}$	
$\mathcal{F}(m)$... differentiable functions on m	$f \in \mathcal{F}(m)$, $f = f(x)$ (round brackets indicate the function dependence)	
$\mathcal{F}(\mathcal{E})$... differentiable functionals on $\mathcal{E}$	$f \in \mathcal{F}(\mathcal{E})$, $f = f[e]$ (square brackets indicate the functional dependence)	
$T_x(m)$... tangent space to m at x	$\tilde{\lambda} \in T_x(m)$ (tilde above the symbol)	$\tilde{\lambda} = \lambda^a \tilde{\partial}_a$
$T_X(\mathcal{M})$... tangent space to $\mathcal{M}$ at X	$\bar{\lambda} \in T_X(\mathcal{M})$ (bar above the symbol)	$\bar{\lambda} = \lambda^\alpha \bar{\partial}_\alpha$
$T_e(\mathcal{E})$... tangent space to $\mathcal{E}$ at e (Sec. 4)	$\mathbf{N} \in T_e(\mathcal{E})$ (boldface)	$\mathbf{N} = \int_{x \in m} \bar{N}(x) \rfloor \delta_{e(x)} = N^{\alpha x} \delta_{\alpha x}$ [Eq. (4.5)]
$T_x^*(m)$... cotangent space to m at x	$\underset{\sim}{\lambda} \in T_x^*(m)$ (tilde below the symbol)	$\underset{\sim}{\lambda} = \lambda_a \underset{\sim}{d} x^a$
$T_X^*(\mathcal{M})$... cotangent space to $\mathcal{M}$ at X	$\underline{\lambda} \in T_X^*(\mathcal{M})$ (bar below the symbol)	$\underline{\lambda} = \lambda_\alpha \underline{d} X^\alpha$
$T_e^*(\mathcal{E})$... cotangent space to $\mathcal{E}$ at e (Sec. 4)	$\mathbf{M} \in T_e^*(\mathcal{E})$ (boldface)	$\mathbf{M} = \int_{x \in m} \underline{M}(x) \llcorner \mathbf{d}e(x) = M_{abc\,\alpha x} \mathbf{d}e^{\alpha x}$ [Eq. (4.20)]
$T_x{}^r{}_s(m)$... space of r-contravariant, s-covariant tensors to m at x	$\tilde{\lambda} \in T_x{}^r{}_s(m)$ (tilde placed over the symbol for contravariant or mixed tensors, below the symbol for covariant tensors)	$\lambda^{a\cdots}{}_{b\cdots}$
$T_X{}^R{}_S(\mathcal{M})$... space of R-contravariant,	$\bar{\lambda} \in T_X{}^R{}_S(\mathcal{M})$ (bar placed over the	$\lambda^{\alpha\cdots}{}_{\beta\cdots}$

S-covariant tensors to $\mathcal{M}$ at X	symbol for contravariant or mixed tensors, below the symbol for covariant tensors)	
$T_{x\underset{\sim}{S}}(m)$... space of s-forms to m at x	$\underset{\sim}{\lambda} \in T_{x\underset{\sim}{S}}(m)$ (tilde below the symbol)	$\lambda_{a\ldots b}$
$T_{X\underset{\sim}{S}}(\mathcal{M})$... space of S-forms to $\mathcal{M}$ at X	$\underline{\lambda} \in T_{X\underset{\sim}{S}}(\mathcal{M})$ (bar below the symbol)	$\lambda_{\alpha\beta\ldots\gamma}$
	For 3-forms to m and 4-forms to $\mathcal{M}$, the form indices are often suppressed and the "densities" notation is used.	
$T_e \overset{R,r}{S,s}(\mathcal{E})$... fiber of e-tensors over $\mathcal{E}$ (Sec. 5)	$\boldsymbol{\lambda} \in T_e \overset{R,r}{S,s}(\mathcal{E})$ (boldface)	$\boldsymbol{\lambda} = \lambda^{\alpha\ldots a\ldots}{}_{\beta\ldots b\ldots}(x)\, \partial_{\alpha\ldots a\ldots}$ [Eq. (5.2)]
$T(m)$, etc. ... tangent bundle at m, etc.		
$\mathcal{T}(m)$, etc. ... (vector) fields over m, etc.		

Moving frames

$e^* = \dfrac{\partial e(x)}{\partial x} = \overline{\underset{\sim}{e}} = e^\alpha_a \underset{\sim}{dx}{}^a \otimes \overline{\partial}_\alpha$... tangent vectors to the embedding e

$\overline{n}$... the unit normal to the embedding

$\{\overline{n}, \overline{\underset{\sim}{e}}\}$... the normal basis in $T_{X=e(x)}(\mathcal{M})$

$\{\mathbf{n}, \mathbf{e}\}$... $\{\overline{n}(x), \overline{\underset{\sim}{e}}(x)\}_{x \in m}$ considered as a normal e-basis in $T_e \overset{1,0}{0,0}(\mathcal{E})$

$\{\delta_{\perp x}, \delta_{\parallel x}\}$ or $\{\delta_{\perp x}, \delta_{ax}\}$... normal $\mathcal{E}$-basis in $T_e(\mathcal{E})$ [Eq. (4.12)]

Delta function

$\delta(x, x') = \delta_{xx'} = \delta_{x\,abc\,x'}$... is considered as a scalar in the first and a scalar density in the second argument. The derivatives $\delta_{,a}(x, x')$ of the δ function are always taken with respect to the first argument.

Metric

$\underset{\sim}{g}$... metric tensor in m

$\underline{g}$... metric tensor in $\mathcal{M}$

$\mathbf{g}_x$... (degenerate) metric tensor in $\mathcal{E}$ [Eqs. (8.13), (8.14)]

$\epsilon = \mp 1$... the indicator of spacetime signature $(\epsilon, 1, 1, 1)$. (Sec. 3)

g... the determinant of $\underset{\sim}{g}$

η... the Levi-Civita 3-form (density)

$\eta_{abc} = g^{1/2}\,\delta_{abc}$, δ_{abc}... the permutation symbol

$\underset{\sim}{K}$... the extrinsic curvature of $e \in \mathcal{E}$

Algebraic operations

$\langle \underset{\sim}{\lambda}, \tilde{\mu} \rangle$... the inner product of a space form $\underset{\sim}{\lambda}$ with a space vector $\tilde{\mu}$

$\overline{\mu} \rfloor \underline{\lambda} = \underline{\lambda} \lfloor \overline{\mu}$... the inner product of a spacetime form $\underline{\lambda}$ with a spacetime vector $\overline{\mu}$

$\perp, \parallel$... normal and tangential projections of spacetime tensors [Eqs. (3.11)–(3.15)]

$\otimes$... direct product

$( )\ [\ ]$... symmetrization and antisymmetrization brackets for a pair of (Latin or Greek) indices: $\lambda_{(ab)} = \frac{1}{2}(\lambda_{ab} + \lambda_{ba})$, $\lambda_{[ab]} = \frac{1}{2}(\lambda_{ab} - \lambda_{ba})$

$(ax \rightleftharpoons bx')$... means: "The same term with the indices a, b and the points x, x' interchanged"

Derivatives

(a) Partial

$\tilde{N} = \tilde{\partial}_{\tilde{N}}$, $\overline{N} = \overline{\partial}_{\overline{N}}$, $\mathbf{N} = \delta_{\mathbf{N}}$... directional derivatives

$\delta_N = \delta_{\epsilon N^x \delta_{\perp x}}$... directional derivative normal to the hypersurface

$\delta_{\tilde{N}} = \delta_{N^{ax} \delta_{ax}}$... directional derivative tangential to the hypersurface

$\delta_{\mathcal{T}}$... directional derivative along a hypersurface tilt (Sec. 10)

$\delta_{\perp}$... directional derivative along a hypersurface translation (Sec. 10)

$,a$ $,\alpha$... partial derivatives with respect to x^a or X^α

(b) Lie brackets and Lie derivatives

$[\ ,\ ]$... Lie bracket between two space or spacetime vectors

$[\ ,\ ]$... Lie bracket between two $\mathcal{E}$-vectors [Eq. (6.1)]

$L_{\tilde{N}}$... Lie derivative of a space tensor field along a space vector field $\tilde{N}$

$\mathbf{L}_{\tilde{N}}$... Lie derivative of an e-tensor field along a space vector field $\tilde{N}$ (Sec. 7)

(c) Covariant derivatives

∇ or $;$... covariant derivative with respect to the Riemannian affine structure in $(\mathcal{M}, \underset{\sim}{g})$

$\Gamma^\alpha{}_{\beta\gamma}(X)$... the Riemannian affine connection in $(\mathcal{M}, \underset{\sim}{g})$

$|$... the covariant derivative with respect to the Riemannian affine structure in $(m, \underset{\sim}{g})$

$\gamma^a{}_{bc}(x)$... the Riemannian affine connection in $(m, \underset{\sim}{g})$

∇... induced symmetrical covariant derivative in hyperspace (Sec. 8)

$\Gamma^{\alpha x}{}_{\beta x'\,\gamma x''} = \Gamma^\alpha{}_{\beta\gamma}(x, x'')\,\delta(x, x')$... the affine connection of ∇ [Eq. (8.5)]

$\overset{*}{\nabla}$... natural asymmetrical covariant derivative in hyperspace (Sec. 2)

$\overset{*}{\Gamma}{}^{\alpha x}{}_{\beta x'\,\gamma x''} = \overset{*}{\Gamma}{}^\alpha{}_\beta(x)_{\gamma x''}\,\delta(x, x')$... the spacetime leg of the affine connection of $\overset{*}{\nabla}$ [Eqs. (11.4), (11.11)]

$\overset{*}{\gamma}{}^{ax}{}_{bx'\,\gamma x''} = \overset{*}{\gamma}{}^a{}_b(x)_{\gamma x''}\,\delta(x, x')$... the space leg of the affine connection of $\overset{*}{\nabla}$ [Eqs. (11.4), (11.6)]

$\Lambda^\alpha{}_{\beta\gamma}(x, x') = \overset{*}{\Gamma}{}^\alpha{}_{\beta\gamma}(x, x') - \Gamma^\alpha{}_{\beta\gamma}(x, x')$... the hyperbitensor characterizing the torsion of $\overset{*}{\nabla}$ [Eq. (11.11)]

$\nabla_N \equiv \nabla_{\mathbf{N}=\epsilon N^x \delta_{\perp x}}$... covariant derivative along a normal deformation $\mathcal{E}$-vector

$\nabla_{\tilde{N}} \equiv \nabla_{\mathbf{N}=N^{ax}\delta_{ax}}$... covariant derivative along a tangential deformation $\mathcal{E}$-vector

$\nabla_{\nearrow}$... covariant derivative along a hypersurface tilt (Sec. 10)

$\nabla_{\perp}$... covariant derivative along a hypersurface translation (Sec. 10)

Integrals

$\int_{x\in m} \underset{\sim}{f}(x) = \int_{x\in m} \underset{\sim}{d}x^a \wedge \underset{\sim}{d}x^b \wedge \underset{\sim}{d}x^c f_{abc}(x)$... integral of a scalar density $\underset{\sim}{f}$ (a 3-form) over m

$\int_{x^a\in \mathbf{R}^3} d^3x$... an ordinary Riemann integral over $\mathbf{R}^3$

$M_{\alpha x} N^{\alpha x}$... integration is implied over a label x repeated in the lower and upper index position (an analog to the summation convention). Here, e.g., $M_{\alpha\, abc}(x)$ may be a spacetime vector-space scalar density, and $N^{\alpha x} \equiv N^\alpha(x)$ a spacetime vector at $X = e(x)$.

3. EMBEDDINGS

The basic concepts of hypersurface field dynamics stem from the theory of embeddings, which we shall review in this section. Let us have an embedding

$$e : x \in m \to X \in \mathcal{M}, \quad X = e(x) \tag{3.1}$$

of a three-dimensional manifold m (space) to a four-dimensional manifold $\mathcal{M}$ (spacetime). The image of m under the embedding e is a hypersurface in spacetime. A definite indentification of points $x \in m$ which are adjacent to the points $X \in \mathcal{M}$ is implied by the term embedding; on the other hand, we have no particular identification of points in mind when we speak about a hypersurface. The hypersurface h may thus be conceived as an equivalence class of embeddings which differ from each other only by a three-dimensional diffeomorphism φ

$$\underset{\circ}{h} \equiv \{e = \underset{\circ}{e} \circ \varphi \mid \varphi \in \mathrm{Diff}(m)\}. \tag{3.2}$$

If we introduce the local coordinates $x^a(x)$ in m and similarly the local coordinates $X^\alpha(X)$ in $\mathcal{M}$, the embedding e becomes locally characterized by four function $e^\alpha(x^a)$ of three coordinates x^a,

$$X^\alpha = e^\alpha(x^a). \tag{3.3}$$

The two sets of functions, $e^\alpha(x^a)$ and $e^\alpha(x^{a'})$,

$$X^\alpha = e^\alpha(x^a) \text{ and } X^\alpha = e^\alpha(x^b(x^{a'})), \tag{3.4}$$

represent the same embedding in two different systems of coordinates.

The mapping $e: x \in m \to e(x) \in \mathcal{M}$ induces the map $e_* = \partial e(x)/\partial x$ of the tangent space $T_x(m)$ into the tangent space $T_{X=e(x)}(\mathcal{M})$. We can consider e_* as a vector in $T_X(\mathcal{M})$ and as a covector in $T_x(m)$, writing

$$e_* = \frac{\partial e(x)}{\partial x} \equiv \overline{\underset{\sim}{e}}. \tag{3.5}$$

Introducing a coordinate basis $\overline{\partial}_\alpha \equiv \partial/\partial X^\alpha$ in $T_X(\mathcal{M})$ and a cobasis of differentials $\underset{\sim}{d}x^a$ in $T^*_x(m)$, we have

$$\overline{\underset{\sim}{e}} = e^\alpha_a \overline{\partial}_\alpha \otimes \underset{\sim}{d}x^a, \quad \text{with } e^\alpha_a \equiv e^\alpha_{,a}. \tag{3.6}$$

If $\mathcal{M}$ is a Riemannian spacetime carrying the metric $\underset{\sim}{g}$, the mapping e induces a metric $\underset{\sim}{g}$ on m, which we interpret as the intrinsic metric of the hypersurface,

$$\underset{\sim}{g} = e^* \underset{\sim}{g} = \underset{\sim}{g}(\overline{\underset{\sim}{e}}, \overline{\underset{\sim}{e}}). \tag{3.7}$$

In a spacetime with the Minkowskian signature $(-,+,+,+)$, we say that the embedding e leads to a spacelike hypersurface h, if the space metric $\underset{\sim}{g}$ is nondegenerate and positive definite [having thus the signature $(+,+,+)$]. In field dynamics, the hypersurfaces are always taken to be spacelike hypersurfaces. However, we also want to know how the indefinite Minkowskian signature reflects itself in the splitted $1+3$ formalism of the hypersurface dynamics. We thus leave undecided whether the signature of $\underset{\sim}{g}$ is $(-,+,+,+)$ or $(+,+,+,+)$, introducing the indicator $\epsilon = \pm 1$, which enables us to treat both signatures $(\epsilon,+,+,+)$ at the same time.

The metric $\underset{\sim}{g}$ is used to bring the spacetime vectors $\overline{\lambda}$ into a one-to-one correspondence with spacetime covectors $\underline{\lambda}$, or, in the coordinate language, to lower the spacetime indices. The metric $\underset{\sim}{g}$ serves the same purpose in the space m. We write

$$\underline{\lambda} = \underset{\sim}{g}(\cdot, \overline{\lambda}), \quad \underline{\tilde{\lambda}} = \underset{\sim}{g}(\cdot, \tilde{\lambda}). \tag{3.8}$$

In particular, we can get different forms $\underline{\tilde{e}}$, $\underline{\underset{\sim}{e}}$, of $\overline{\underset{\sim}{e}}$. Applying a similar convention to tensors of higher order, we write $\overline{g}$ and $\tilde{g}$ for the contravariant metric tensors in $\mathcal{M}$ and m, respectively. Such notation, however, becomes impractical for higher order mixed tensors. We shall write $^-$ and $^\sim$ in the upper position in the case of mixed tensors; to keep track of their detailed character, the component notation is much more convenient than the abstract notation.

The image $e_*(T_x(m))$ of the tangent space $T_x(m)$ is a three-dimensional vector subspace of $T_{X=e(x)}(\mathcal{M})$. A normal $\underline{n}$ to the hypersurface is a nonzero covector $\underline{n} \in T^*_{X=e(x)}(\mathcal{M})$ which is orthogonal to all tangent vectors in $e_*(T_x(m))$,

$$\overline{\underset{\sim}{e}} \rfloor \underline{n} = 0. \tag{3.9}$$

We can normalize $\underline{n}$ so that

$$\overline{g}(\underline{n}, \underline{n}) = \epsilon; \tag{3.10}$$

in spacetime, the normal to a spacelike hypersurface is a timelike vector.

The spacetime vectors $\overline{n}$ and $\overline{\underset{\sim}{e}}$ together form a basis in $T_{X=e(x)}(\mathcal{M})$. An arbitrary spacetime vector $\overline{\lambda}$ may be decomposed with respect to this basis as

$$\overline{\lambda} = \lambda^\perp \overline{n} + \langle \overline{\underset{\sim}{e}}, \tilde{\lambda}^{\parallel} \rangle. \tag{3.11}$$

The components $\lambda^\perp$ and $\tilde{\lambda}^{\parallel}$ are a space scalar and a space vector, respectively,

$$\lambda^\perp = \epsilon \overline{\lambda} \rfloor \underline{n} = \epsilon \underset{\sim}{g}(\overline{\lambda}, \overline{n}), \quad \tilde{\lambda}^{\parallel} = \overline{\lambda} \rfloor \underline{\tilde{e}} = \underset{\sim}{g}(\overline{\lambda}, \tilde{\underset{\sim}{e}}). \tag{3.12}$$

In terms of local coordinates, $\tilde{\lambda}^{\parallel} = \lambda^a \tilde{\partial}_a$, and Eqs. (3.11) and (3.12) read

$$\lambda^\alpha = \lambda^\perp n^\alpha + \lambda^a e^\alpha_a, \quad \lambda^\perp = \epsilon \lambda^\alpha n_\alpha, \quad \lambda^a = \lambda^\alpha e^a_\alpha. \tag{3.13}$$

Similarly, we can decompose an arbitrary spacetime tensor. In general, the abstract notation becomes more and more cumbersome for higher order tensors, and it is then more convenient to use the local coordinates for bookkeeping purposes. As an example, we write down the decomposition formulas for a covariant second order spacetime tensor:

$$\lambda_{\alpha\beta} = \lambda_{\perp\perp} n_\alpha n_\beta + \lambda_{\perp b} n_\alpha e^b_\beta + \lambda_{a\perp} e^a_\alpha n_\beta + \lambda_{ab} e^a_\alpha e^b_\beta, \tag{3.14}$$

with

$$\begin{aligned} &\lambda_{\perp\perp} = \lambda_{\alpha\beta} n^\alpha n^\beta, \quad \lambda_{\perp b} = \epsilon \lambda_{\alpha\beta} n^\alpha e^\beta_b, \\ &\lambda_{a\perp} = \epsilon \lambda_{\alpha\beta} e^\alpha_a n^\beta, \quad \lambda_{ab} = \lambda_{\alpha\beta} e^\alpha_a e^\beta_b. \end{aligned} \tag{3.15}$$

Equations (3.15) exemplify the general rule how to use the indicator ϵ when calculating the space components of tensors: The formula for the component contains the indicator when it carries the $\perp$ projection an odd number of times.

Applying Eqs. (3.14) and (3.15) to the metric tensor itself, we get the completeness relation for the cobasis $\{\underline{n}, \underline{e}\}$,

$$g_{\alpha\beta} = g_{ab} e^a_\alpha e^b_\beta + \epsilon n_\alpha n_\beta. \tag{3.16}$$

Together with the metric, the Riemannian affine structure is induced in m by the embedding. To get the covariant derivative $\tilde{\lambda}|_{\tilde{N}}$ of a vector field $\tilde{\lambda} \in T(m)$ along the vector $\tilde{N} \in T_x(m)$, one projects $\tilde{\lambda}$ and $\tilde{N}$ into $\overline{\lambda} = e_*(\tilde{\lambda}) \in T(\mathcal{M})$ and $\overline{N} = e_*(\tilde{N}) \in T_{X=e(e)}(\mathcal{M})$, takes the covariant derivative $\overline{\lambda}_{;\overline{N}}$ with respect to the spacetime Riemannian structure $(\mathcal{M}, \underset{\sim}{g})$, and projects the resulting vector back into $T_x(m)$. In terms of local coordinates,

$$\lambda^a{}_{|b} = e^a_\alpha (\lambda^c e^\alpha_c)_{;\beta} e^\beta_b. \tag{3.17}$$

From Eq. (3.17), one gets the affine connection $\gamma^a{}_{bc}$ on m as

$$\gamma^a{}_{bc} = e^a_\alpha e^\alpha_{b;\gamma} e^\gamma_c. \tag{3.18}$$

One can check that the affine connection (3.18) is the Riemannian affine connection associated with the metric $\underset{\sim}{g}$, so that

$$\underset{\sim}{g}_{|\tilde{N}} = 0 \qquad \forall \tilde{N}. \tag{3.19}$$

The covariant derivative $\underline{n}_{;\overline{N}}$ of the normal $\underline{n}$ along a tangent vector $\overline{N} = \langle \overline{\underset{\sim}{e}}, \tilde{N} \rangle$ to the hypersurface measures the bending of this hypersurface in the embedding spacetime $(\mathcal{M}, \underset{\sim}{g})$. Because the magnitude of $\underline{n}$ is everywhere the same, $\overline{n} \rfloor \underline{n} = \epsilon$, the normal component of the spacetime covector $\underline{n}_{;\overline{N}}$ vanishes,

$$\overline{n} \rfloor \underline{n}_{;\overline{N}} = 0, \quad \text{with } \overline{N} = \langle \overline{\underset{\sim}{e}}, \tilde{N} \rangle. \tag{3.20}$$

The surviving tangential components

$$\overline{M} \rfloor \underline{n}_{;\tilde{N}}, \quad \text{with } \overline{M} = \langle \overline{\underset{\sim}{e}}, \tilde{M} \rangle, \tag{3.21}$$

lead, due to the linearity of $\rfloor$ and $;\overline{N}$ operations, to the extrinsic curvature tensor

$$\underset{\sim}{K}(\tilde{M}, \tilde{N}) \equiv - \overline{M} \rfloor \underline{n}_{;\tilde{N}}. \tag{3.22}$$

In terms of local coordinates, Eqs. (3.20) and (3.22) read

$$0 = n^\alpha n_{\alpha;b}, \quad K_{ab} = - e^\alpha_a n_{\alpha;b}, \tag{3.23}$$

where $;b$ is the spacetime covariant derivative along the x^b coordinate line. From here,

$$n_{\alpha;b} = - K_{ab} e^a_\alpha. \tag{3.24}$$

Because e^α_a and n_α are orthogonal, $n_\alpha e^\alpha_a = 0$, the ex-

trinsic curvature can also be written as

$$K_{ab} = n_\alpha e^\alpha_{a;b}. \tag{3.25}$$

The relation

$$e^\alpha_{a;b} = e^\alpha_{b;a} \tag{3.26}$$

then ensures the symmetry of K_{ab}.

As a pendant to Eq. (3.24), one can get the covariant derivative $e^\alpha_{a;b}$. Equation (3.25) gives the normal component of this spacetime vector, while Eq. (3.18) gives its tangential component. Therefore,

$$e^\alpha_{a;b} = \epsilon K_{ab} n^\alpha + \gamma^c{}_{ab} e^\alpha_c. \tag{3.27}$$

The last equation is called the Gauss–Weingaarten equation. Equations (3.24) and (3.27) tell us how the basis $\{\bar{n}, \underline{e}\}$ changes if we go along the hypersurface.

4. THE SPACE OF EMBEDDINGS AND HYPERSPACE

Heuristically, at least, all embeddings e which lead to a spacelike hypersurface in $(\mathcal{M}, \underline{g})$ form themselves an infinitely dimensional manifold $\mathcal{E}$,

$$\mathcal{E} \equiv \{e \,|\, \bar{g}(\underline{n}, \underline{n}) = \epsilon\}, \tag{4.1}$$

which we shall call the *space of embeddings*. Any choice of local coordinates in m and $\mathcal{M}$ induces a coordinate map in $\mathcal{E}$; indeed, any four functions (3.3) restricted by the condition that

$$g_{ab}(x^c) \equiv g_{\alpha\beta}(e^\gamma(x^c))\, e^\alpha_a(x^c)\, e^\beta_b(x^c) \tag{4.2}$$

is a regular positive definite matrix, may be considered as local coordinates of a point $e \in \mathcal{E}$.

Similarly, we shall treat the collection of all hypersurfaces (3.2) as an infinitely dimensional manifold

$$\mathcal{H} \equiv \mathcal{E}/\mathrm{Diff}(m), \tag{4.3}$$

which we shall call *hyperspace*. Hyperspace plays the same basic role for the dynamics of tensor fields in spacetime as the time manifold does for the dynamics of ordinary particles. A single hypersurface h is the proper relativistic generalization of the concept of "an instant of time." Technically, it is easier to represent hypersurfaces by embeddings, and ensure that all operations behave properly under the three-dimensional diffeomorphisms. Let us thus start by studying the differential geometry in the space of embeddings.

A one-parameter family of embeddings,

$$e(t) : t \in \mathbf{R}, \quad x \in m \to X = e(t, x) \in \mathcal{M},$$

i.e., a curve in $\mathcal{E}$, represents a continuous deformation of a hypersurface in spacetime. The tangent vectors

$$\mathbf{N} = \left.\frac{de(t)}{dt}\right|_{t=t_0} \tag{4.4}$$

to all such curves passing through the same point $e \equiv e(t_0)$ in $\mathcal{E}$ fill the tangent space $T_e(\mathcal{E})$. We shall call the vectors $\mathbf{N}$ from $T_e(\mathcal{E})$ the $\mathcal{E}$-vectors, to distinguish them from the spacetime vectors $\bar{N}$ and the space vectors $\tilde{N}$. A tangent vector in $T_e(\mathcal{E})$ can be considered as a linear differential operator acting on the ring $\mathcal{F}(\mathcal{E})$ of differentiable functionals of the embedding e. Indeed, if $f[e] \in \mathcal{F}(\mathcal{E})$ is such a functional, then

$$\mathbf{N}f \equiv \left.\frac{df[e(t,x)]}{dt}\right|_{t=t_0} = \int_{x\in m} \bar{N}(x) \,\lrcorner\, \delta_{e(x)} f[e(x)]. \tag{4.5}$$

Here,

$$\bar{N}(x) \equiv \left.\frac{\partial e(t,x)}{\partial t}\right|_{t=t_0} \tag{4.6}$$

is a spacetime vector field along the embedding e, i.e., the mapping

$$\bar{N}(x) : x \in m \to \bar{N}(x) \in T_{X=e(x)}(\mathcal{M}), \tag{4.7}$$

and the variational derivative

$$\delta_{e(x)} \equiv \frac{\delta}{\delta e(x)} \tag{4.8}$$

is a field of spacetime vector—space density valued linear differential operators, acting on $\mathcal{F}(\mathcal{E})$ and defined along the embedding e, i.e., $\delta_{e(x)}$ is the mapping

$$\delta_{e(x)} : f \in \mathcal{F}(\mathcal{E}) \to \delta_{e(x)} f \in T_e {}^{1,0}_{0,3}(\mathcal{E})$$

(see Secs. 2 and 5 for the notation). After the local coordinates (3.3) are introduced in $\mathcal{E}$, the functionals f become represented by some functionals $f[e^\alpha(x^a)]$ of $e^\alpha(x^a)$, and the variational derivative $\delta_{e(x)}$ may be expressed as

$$\delta_{e(x)} = \underline{d}X^\alpha \delta_{\alpha x}, \quad \delta_{\alpha x} = \frac{\delta}{\delta e^\alpha(x^a)}. \tag{4.9}$$

Equation (4.5) then assumes the familiar form

$$\left.\frac{d}{dt} f[e^\alpha(t, x^a)]\right|_{t=t_0} = \int_{x^a \in \mathbf{R}^3} d^3x \left.\frac{\partial e(t, x^a)}{\partial t}\right|_{t=t_0} \left.\frac{\delta f[e^\beta(x^b)]}{\delta e^\alpha(x^a)}\right|_{e^\alpha(t_0, x^a)}. \tag{4.10}$$

Equation (4.5) tells us that the variational derivatives (4.8) form a "coordinate basis" in $T_e(\mathcal{E})$ and the spacetime vector field $\bar{N}(x)$ can be considered as the component expression of an $\mathcal{E}$-vector $\mathbf{N}$ with respect to this basis. We shall refer to $\delta_{e(x)}$ as the "coordinate $\mathcal{E}$-basis" in $T_e(\mathcal{E})$. More explicitly, after the local coordinates (3.3) are actually introduced into $\mathcal{E}$, the coordinate $\mathcal{E}$-basis may be represented by the variational derivatives $\delta/\delta e^\alpha(x^a)$, and the components $\bar{N}(x)$ of $\mathbf{N}$ by the functions $N^\alpha(x^a)$. The relation between the "abstract" coordinate $\mathcal{E}$-basis $\delta_{e(x)}$ and its actual coordinate expression $\delta/\delta e^\alpha(x^a)$ is given by Eq. (4.9).

From the coordinate $\mathcal{E}$-basis (4.9) one can pass to an arbitrary "moving frames" $\mathcal{E}$-basis in $T_e(\mathcal{E})$. Of particular interest is the normal $\mathcal{E}$-basis $\{\delta_{\perp x}, \delta_{\|x}\}$, determined by the orthogonal decomposition of the operator covector densities (4.8) according to the scheme

$$\underline{\delta}_{e(x)} = \underline{n}(x)\, \delta_{\perp x} + \langle \underline{\tilde{e}}, \underline{\delta}_{\|x} \rangle. \tag{4.11}$$

The expressions

$$\delta_{\perp x} \equiv \epsilon \underline{n}(x) \,\lrcorner\, \underline{\delta}_{e(x)}, \quad \underline{\delta}_{\|x} \equiv \underline{\tilde{e}}(x) \,\lrcorner\, \delta_{e(x)} \tag{4.12}$$

are a space scalar density and a space vector density valued linear differential operators acting on $\mathcal{F}(\mathcal{E})$ and defined along the embedding e. We have underlined the symbols $\underline{\delta}_{e(x)}$ and twiddled the symbols $\underline{\delta}_{\|x}$, to stress an appropriate character of the quantities $\underline{\delta}_{e(x)} f[e(x)]$ and $\delta_{\|x} f[e(x)]$.

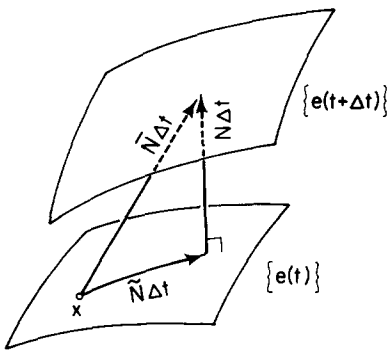

FIG. 1. Lapse-shift decomposition. The deformation $\mathcal{E}$-vector **N** is decomposed with respect to the normal $\mathcal{E}$-basis into the component lapse function $N(x)$ and shift vector $\tilde{N}(x)$.

The operators (4.12) generate the deformations of the embedding which are perpendicular and parallel to the hypersurface from the point of view of the embedding spacetime $(\mathcal{M}, \underset{\sim}{g})$. The deformation $\mathcal{E}$-vector **N** may be decomposed with respect to the normal $\mathcal{E}$-basis (4.12) as

$$\mathbf{N} = \int_{x\in m} (\epsilon N(x)\, \delta_{\perp x} + N^a(x)\, \delta_{ax}). \tag{4.13}$$

The components $N(x)$, $\tilde{N}(x)$ are related to the components $\bar{N}(x)$ of **N** in the coordinate $\mathcal{E}$-basis by the recipe (3.11), (3.12),

$$\bar{N} = N\bar{n} + \langle \underset{\sim}{\bar{e}}, \tilde{N}\rangle, \tag{4.14}$$

$$N = \epsilon \bar{N} \rfloor \underline{n}, \quad \tilde{N} = \bar{N} \rfloor \underset{\sim}{\bar{e}}. \tag{4.15}$$

Geometrically, $N\Delta t$ is the proper time which an observer moving perpendicular to the hypersurface $\{e(t)\}$ needs in order to reach the neighboring hypersurface $\{e(t+\Delta t)\}$, and $\tilde{N}\Delta t$ shows how far he must first go from the point x in the direction $\tilde{N}$ along the embedding $e(t)$ in order to land at the point x of the embedding $e(t+\Delta t)$, if he lauches into spacetime perpendicular to the first hypersurface (Fig. 1). The scalar field $N \in \mathcal{F}(m)$ and the vector field $\tilde{N} \in \overline{\mathcal{T}}(m)$ are called the lapse function and the shift vector, respectively.

Among all $\mathcal{E}$-vectors, we can select those which are intersections of a spacetime vector field $\bar{N}(X)$ by the hypersurface

$$\mathbf{N} = \int_{x\in m} \bar{N}(e(x)) \rfloor \underline{\delta}_{e(x)}. \tag{4.16}$$

We shall call them spacetime $\mathcal{E}$-vectors. Not all $\mathcal{E}$-vectors, of course, are of this kind. Notable counterexamples are the normal and tangential $\mathcal{E}$-vectors, generated by the spacetime fields $\bar{n}(x)$ and $\underset{\sim}{\bar{e}}(x)$ defined only along the embedding e, not in the whole spactime $\mathcal{M}$,

$$\begin{aligned} \mathbf{n} &= \int_{x\in m} \bar{n}(x) \rfloor \underline{\delta}_{e(x)} = \epsilon \int_{x\in m} \delta_{\perp x}, \\ \mathbf{e} &= \int_{x\in m} \underset{\sim}{\bar{e}}(x) \rfloor \underline{\delta}_{e(x)} = \int_{x\in m} \underset{\sim}{\delta}_{\|x}. \end{aligned} \tag{4.17}$$

The operations which we shall further consider, like the Lie bracket between the $\mathcal{E}$-vectors, or the covariant derivative of an $\mathcal{E}$-vector, simplify when the $\mathcal{E}$-vectors are spacetime $\mathcal{E}$-vectors.

The dual space to $T_e(\mathcal{E})$ is the space $T_e^*(\mathcal{E})$ of $\mathcal{E}$-covectors. An $\mathcal{E}$-covector is a linear functional acting on the space of $\mathcal{E}$-vectors $T_e(\mathcal{E})$,

$$\mathbf{M}: \mathbf{N} \in T_e(e) \to \langle \mathbf{M}, \mathbf{N}\rangle \in \mathbf{R}. \tag{4.18}$$

Introduce the cobasis of $\mathcal{E}$-differentials, $\mathrm{d}e(x)$, which is dual to the coordinate $\mathcal{E}$-basis $\delta_{e(x)}$,

$$\langle \mathrm{d}e^\alpha(x),\ \delta_{\beta x'}\rangle = \delta^\alpha_\beta \delta(x, x'), \tag{4.19}$$

and decompose **M** with respect to this cobasis, getting

$$\begin{aligned} \mathbf{M} &= \int_{x\in m} \underline{M}(x) \llcorner \mathrm{d}e(x) = \int_{x\in m} M_\alpha(x)\, \mathrm{d}e^\alpha(x) \\ &= \int_{x^a \in \mathbf{R}^3} \underset{\sim}{d}x^a \wedge \underset{\sim}{d}x^b \wedge \underset{\sim}{d}x^c\, M_{abc\alpha}(x^d)\, \mathrm{d}e^\alpha(x^d) \\ &\equiv M_{\alpha x}\, \mathrm{d}e^{\alpha x}. \end{aligned} \tag{4.20}$$

The component expression of **M** is thus a field of spacetime covectors—space densities defined along the embedding. Using Eqs. (4.19) and (4.20), we get the coordinate expression of the $\langle\ \rangle$ product,

$$\begin{aligned} \langle \mathbf{M}, \mathbf{N}\rangle &= \int_{x^a \in \mathbf{R}^3} \underset{\sim}{d}x^a \wedge \underset{\sim}{d}x^b \wedge \underset{\sim}{d}x^c\, M_{abc\alpha}(x^d)\, N^\alpha(x^d) \\ &\equiv M_{\alpha x} N^{\alpha x}. \end{aligned} \tag{4.21}$$

One can finally introduce $\mathcal{E}$-tensors of an arbitrary rank as multilinear functionals acting on the direct product of a number of tangent and cotangent spaces $T_e(\mathcal{E})$ and $T_e^*(\mathcal{E})$. It is hardly necessary to go into details of this well-known procedure. The elements of $\mathcal{F}(\mathcal{E})$ can also be considered as $\mathcal{E}$-scalars.

5. *e*-TENSORS

$\mathcal{E}$-tensors are multilocal objects on m, their components being defined over $m \otimes m \otimes \cdots \otimes m$. For example, a second rank contravariant $\mathcal{E}$-tensor **P** is characterized by a bivector $P^{\alpha x\, \beta x'}$,

$$\mathbf{P} = P^{\alpha x\, \beta x'}\, \delta_{\alpha x} \otimes \delta_{\beta x'}.$$

In field dynamics, however, we are mostly dealing with spacetime tensor fields defined along the embeddings and with the normal and tangential projections of these fields to the hypersurface. Such fields are local objects on m. We thus introduce what we shall call the *bundle of e-tensors over* $\mathcal{E}$, denoted by $T^{R,r}_{S,s}(\mathcal{E})$. An element of the fiber $T_e{}^{R,r}_{S,s}(\mathcal{E})$ at e is a field of mixed spacetime–space tensors along e, the spacetime tensor rank being $\binom{R}{S}$ and the space tensor rank being $\binom{r}{s}$. From the basis fields $\bar{\partial}_a(x)$, $\underset{\sim}{d}x^a(x)$, and $\bar{\partial}_\alpha(e(x))$, $\underset{\sim}{d}X^\alpha(e(x))$ along e, we form an e-basis in $T_e{}^{R,r}_{S,s}(\mathcal{E})$ as their direct product taken point by point in m:

$$\begin{aligned} &\partial_{\alpha\cdots a\cdots}{}^{\beta\cdots b\cdots}(x) \\ &= \underbrace{\bar{\partial}_\alpha(e(x)) \otimes \cdots}_{R}\ \underbrace{\tilde{\partial}_a(x) \otimes \cdots}_{r}\ \underbrace{\otimes\, \underset{\sim}{d}X^\beta(e(x)) \otimes \cdots}_{S}\ \underbrace{\underset{\sim}{d}x^b(x) \otimes \cdots}_{s} \end{aligned} \tag{5.1}$$

Thus, $\boldsymbol{\lambda} \in T_e{}^{R,r}_{S,s}(\mathcal{E})$ may be expressed in the form

$$\boldsymbol{\lambda} = \lambda^{\overbrace{\alpha\cdots}^{R}\ \overbrace{a\cdots}^{r}}_{\underbrace{\beta\cdots}_{S}\ \underbrace{b\cdots}_{s}}(x)[e]\, \partial_{\alpha\cdots a\cdots}{}^{\beta\cdots b\cdots}(x). \tag{5.2}$$

Among the natural algebraic operations on e-tensors, let us mention the direct product $\otimes$ of two e-tensors, the lowering and raising of the spacetime and space indices by the metric tensors $\underline{g}$ and $\underset{\sim}{g}$, the spacetime and space contractions, the projections of spacetime

indices into $\perp$ and $\parallel$ directions to the hypersurface, and finally the lifting of a space tensor index into a spacetime tensor index according to the rule

$$\lambda^{a\cdots}{}_{\ldots}(x) \to \lambda^{\alpha\cdots}{}_{\ldots}(x) = e^{\alpha}_{a}(x)\,\lambda^{a\cdots}{}_{\ldots}(x). \tag{5.3}$$

All these operations are local, taking place point by point over the manifold

6. LIE BRACKETS IN $\mathcal{E}$

If **M** and **N** are two $\mathcal{E}$-vector fields, we can define their Lie bracket $[\mathbf{M}, \mathbf{N}]$ in the standard way,

$$[\mathbf{M}, \mathbf{N}] = \mathbf{MN} - \mathbf{NM}. \tag{6.1}$$

$[\mathbf{M}, \mathbf{N}]$ is again an $\mathcal{E}$-vector field. Its components with respect to the coordinate $\mathcal{E}$-basis $\delta_{\alpha x}$ are

$$[\mathbf{M}, \mathbf{N}]^{\alpha x} = M^{\beta x'}\,\delta_{\beta x'} N^{\alpha x} - N^{\beta x'}\,\delta_{\beta x'} M^{\alpha x}. \tag{6.2}$$

In particular, when **M** and **N** are intersections of two spacetime vector fields, $\bar{M}(X)$ and $\bar{N}(X)$, by the embedding e, Eq. (6.2) gives the relation between the Lie bracket [ ] in the space of embeddings and the Lie bracket [ ] in spacetime,

$$[\mathbf{M}, \mathbf{N}]^{\alpha x} = [\bar{M}, \bar{N}]^{\alpha}\,|_{X^{\beta}=e^{\beta}(x)}. \tag{6.3}$$

The Lie bracket between any two $\mathcal{E}$-vectors $\delta_{\alpha x}$ vanishes,

$$[\delta_{\alpha x}, \delta_{\beta x'}] = 0. \tag{6.4}$$

This expresses the fact that $\delta_{\alpha x}$ is a coordinate $\mathcal{E}$-basis. On the other hand, the Lie brackets between the $\mathcal{E}$-vectors of the normal $\mathcal{E}$-basis $\{\delta_{\perp x}, \delta_{\alpha x}\}$ are different from zero. They play an important role in hypersurface dynamics and we thus proceed with their evaluation.

The components of $\delta_{\perp x'}$ and $\delta_{\alpha x'}$ in the coordinate $\mathcal{E}$-basis are

$$[\delta_{\perp x'}]^{\alpha x} = \epsilon n^{\alpha}(x)\,\delta(x, x'), \quad [\delta_{\alpha x'}]^{\alpha x} = e^{\alpha}_{a}(x)\,\delta(x, x'). \tag{6.5}$$

Subsituting them into Eq. (6.2), we get

$$[\delta_{\alpha x'}, \delta_{b x''}]^{\alpha x} = e^{\beta}_{a}(x')\,\delta_{\beta x'}\,e^{\alpha}_{b}(x'')\,\delta(x, x'') - (a x' \leftrightarrow b x''), \tag{6.6}$$

$$[\delta_{\alpha x'}, \delta_{\perp x''}]^{\alpha x} = \epsilon\, e^{\beta}_{a}(x')\,\delta_{\beta x'} n^{\alpha}(x'')\,\delta(x, x'') - \epsilon n^{\beta}(x'')\,\delta_{\beta x''} e^{\alpha}_{a}(x')\,\delta(x, x'), \tag{6.7}$$

$$[\delta_{\perp x'}, \delta_{\perp x''}]^{\alpha x} = n^{\beta}(x')\,\delta_{\beta x'} n^{\alpha}(x)\,\delta(x, x'') - (x' \leftrightarrow x''). \tag{6.8}$$

The variational derivative of $e^{\alpha}_{a}(x) = e^{\alpha}{}_{,a}(x)$ is given by

$$\delta_{\beta x'} e^{\alpha}_{a}(x) = \delta^{\alpha}_{\beta}\,\delta_{,a}(x, x'). \tag{6.9}$$

First, substitute Eq. (6.9) into Eq. (6.6). An important identity

$$f(x')\,\delta_{,a}(x, x')\,g(x) = f(x)\,\delta_{,a}(x, x')\,g(x) + f_{,a}(x)\,\delta(x, x')\,g(x) \tag{6.10}$$

enables us to evaluate all coefficients of $\delta_{,a}(x, x')$ at the same point x. Equation (6.6) then gives

$$-[\delta_{\alpha x'}, \delta_{b x''}] = \delta_{,a'}(x', x'')\,\delta_{b x'} - (a x' \leftrightarrow b x''). \tag{6.11}$$

Second, substitute Eq. (6.9) into Eq. (6.7) and again use the identity (6.10). The last term in Eq. (6.7) then becomes

$$-\epsilon n^{\alpha}(x')\,\delta_{,a'}(x', x'')\,\delta(x, x') - \epsilon n^{\alpha}{}_{,a}(x')\,\delta(x', x'')\,\delta(x, x'). \tag{6.12}$$

The first term on the right-hand side of Eq. (6.7) just cancels the last term in the expression (6.12), because

$$e^{\beta}_{a}(x')\,\delta_{\beta x'} n^{\alpha}(x'') = n^{\alpha}{}_{,a}(x'')\,\delta(x'', x'). \tag{6.13}$$

These rearrangements bring Eq. (6.7) into the final form

$$-[\delta_{\alpha x'}, \delta_{\perp x''}] = \delta_{,a'}(x', x'')\,\delta_{\perp x'}. \tag{6.14}$$

Third, turn to Eq. (6.8). The variational derivative of $n^{\alpha}(x)$ is obtained indirectly, by varying the equations

$$g_{\alpha\beta} n^{\alpha} n^{\beta} = \epsilon, \quad g_{\alpha\beta} e^{\alpha}_{a} n^{\beta} = 0 \tag{6.15}$$

and using the variational formula (6.9). We get

$$\delta_{\beta x'} n^{\alpha}(x) = -e^{a\alpha}(x)\,n_{\beta}(x)\,\delta_{,a}(x, x') + A^{\alpha}_{\beta}(x)\,\delta(x, x'). \tag{6.16}$$

The detailed structure of $A^{\alpha}_{\beta}(x)$ need not interest us, because this term drops out under the interchange $(x \leftrightarrow x'')$. Using again the identity (6.10), we cast Eq. (6.8) into the form

$$-[\delta_{\perp x'}, \delta_{\perp x''}] = -\epsilon g^{ab}(x')\,\delta_{,a'}(x', x'')\,\delta_{b x'} - (x' \leftrightarrow x''). \tag{6.17}$$

Equations (6.11), (6.14), and (6.17) are the closing relations for the normal $\mathcal{E}$-basis under the Lie bracket operation [ ]. As explained in Ref. 7, pure geometrodynamics may be reconstructed as the unique "representation" of these closing relations.

Equations (6.11), (6.14), and (6.17) enable us to find the components of the Lie bracket between two $\mathcal{E}$-vectors **M** and **N** in the normal $\mathcal{E}$-basis. Decomposing each of these $\mathcal{E}$-vectors with respect to this basis,

$$\mathbf{M} = \epsilon M^{\perp x}\,\delta_{\perp x} + M^{a x}\,\delta_{a x}, \quad \mathbf{N} = \cdots, \tag{6.18}$$

and using the closing relations (6.11), (6.14), (6.17), we get

$$[\mathbf{M}, \mathbf{N}]^{\perp x} = \mathbf{M} N^{\perp}(x) - \mathbf{N} M^{\perp}(x) + (M^{\perp}{}_{,a}(x)\,N^{a}(x) - N^{\perp}{}_{,a}(x)\,M^{a}(x)), \tag{6.19}$$

$$[\mathbf{M}, \mathbf{N}]^{a x} = \mathbf{M} N^{a}(x) - \mathbf{N} M^{a}(x) + \epsilon(M^{\perp} N^{\perp,a} N^{\perp} - M^{\perp,a} N^{\perp}) - [\tilde{M}, \tilde{N}]^{a}. \tag{6.20}$$

7. $L_{\tilde{N}}$ DERIVATIVE AND HYPERTENSORS

Besides the Lie bracket of two $\mathcal{E}$-vector fields, another Lie derivative type operation plays an important role in hypersurface dynamics, telling us when an e-tensor field can be interpreted as being defined over hyperspace rather than over the space of embeddings. We shall call this operation the Lie derivative $\mathbf{L}_{\tilde{N}}$ of an e-tensor field λ along a space vector field $\tilde{N}$. We arrive at it by studying the behavior of λ under a one-parameter family φ_t of space diffeomorphisms. The diffeomorphism $\varphi \in \mathrm{Diff}(m)$ naturally induces a diffeomorphism $e \to e \circ \phi$ in $\mathcal{E}$, which can be interpreted as a tangential displacement of the hypersurface e (Fig. 2).

From the field $\lambda \in \mathcal{T}^{R,r}_{S,s}(\mathcal{E})$, take an e-tensor $\lambda[e] \in T_e{}^{R,r}_{S,s}(\mathcal{E})$ at the embedding e. Recall that $\lambda[e]$ represents a field of mixed space—spacetime tensors

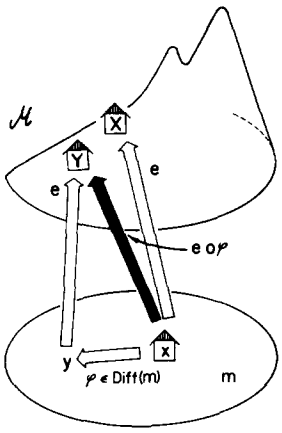

FIG. 2. Tangential displacement. The space diffeomorphism $\varphi \in \mathrm{Diff}(m)$ induces a diffeomorphism $e \circ \varphi \in \mathrm{Diff}(\mathcal{E})$; it can be interpreted as a tangential displacement of the embedding which leaves the hypersurface fixed in spacetime.

along e; for each $x \in m$, the space leg of $\lambda(x)[e]$ is standing in $T^r_{xs}(m)$ and the spacetime leg of $\lambda(x)[e]$ is standing in $T_{e(x)}\,{}^R_S(\mathcal{M})$. Map now $\lambda[e]$, i.e., the field $\lambda(x)[e]$, by φ_t into

$$\varphi_t(\lambda[e]) = \underbrace{\varphi_t^{*-1}}\otimes\cdots\otimes\underbrace{\varphi_t^{*}}\otimes\cdots\lambda(\varphi_t(x))[e],$$

where the inverse mapping φ_t^{*-1} applies to the contravariant and the direct mapping φ_t^* to the covariant space components of λ, while the spacetime components are left untouched by φ. The space leg of $\varphi_t(\lambda[e])$ is standing in $T^r_{xs}(m)$, whereas the spacetime leg of $\varphi_t(\lambda[e])$ is standing in $T_{e\circ\varphi_t(x)}\,{}^R_S(\mathcal{M})$. In other words, $\varphi_t(\lambda[e])$ is an e-tensor in $T_{e\circ\varphi_t}\,{}^{R,r}_{S,s}(\mathcal{E})$. To get an operation which maps $T^{R,r}_{S,s}(\mathcal{E})$ into itself, subtract from $\varphi_t(\lambda[e])$ that element $\lambda[e\circ\varphi_t] \in T_{e\circ\varphi_t}\,{}^{R,r}_{S,s}(\mathcal{E})$ from $T^{R,r}_{S,s}(\mathcal{E})$ which lies at $e\circ\varphi_t$. We thus define the derivative $\mathbf{L}_{\tilde N}\lambda$ of λ with respect to $\tilde N(x) = d\varphi_t/dt\,|_{t=0}$ by the formula

$$\mathbf{L}_{\tilde N}\lambda = \lim_{t\to 0}\frac{1}{t}\{\underbrace{\varphi_t^{*-1}\otimes\cdots\otimes}_{r}\underbrace{\varphi_t^{*}\otimes\cdots}_{s}\lambda(\varphi_t(x))[e] - \lambda(x)[e\circ\varphi_t]\}. \tag{7.1}$$

In the coordinate e-basis,

$$\begin{aligned}\mathbf{L}_{\tilde N}\,&\lambda^{\alpha\cdots a\cdots}{}_{\beta\cdots b\cdots}(x)\\ &= N^c(x)\,\tilde\partial_c\lambda^{\alpha\cdots a\cdots}{}_{\beta\cdots b\cdots}(x)\\ &\quad - (N^c e_c^\gamma)^{x'}\,\delta_{\gamma x'}\lambda^{\alpha\cdots a\cdots}{}_{\beta\cdots b\cdots}(x)\\ &\quad - \lambda^{\alpha\cdots c\cdots}{}_{\beta\cdots b\cdots}(x)\,\tilde\partial_c N^a(x) - \cdots\\ &\quad + \lambda^{\alpha\cdots a\cdots}{}_{\beta\cdots c\cdots}(x)\,\tilde\partial_b N^c(x).\end{aligned} \tag{7.2}$$

The Lie derivatives $\mathbf{L}_{\tilde N}$ and $L_{\tilde N}$ are connected by the formula

$$L_{\tilde N} = L_{\tilde N} - (N^c e_c^\gamma)^{x'}\,\delta_{\gamma x'} = L_{\tilde N} - \delta_{\tilde N}.$$

Note that the Lie derivative $L_{\tilde N}$ does not produce a spacetime tensor field along the embedding when applied to a spacetime tensor field, whereas the Lie derivative $\mathbf{L}_{\tilde N}$ does.

The Lie derivative $\mathbf{L}_{\tilde N}$ is a derivation on the algebra of e-tensors:

$$\begin{aligned}&\mathbf{L}_{\tilde N}(\lambda+\mu) = \mathbf{L}_{\tilde N}\lambda + \mathbf{L}_{\tilde N}\mu,\\ &\mathbf{L}_{\tilde N}(\lambda\otimes\mu) = \mathbf{L}_{\tilde N}\lambda\otimes\mu + \lambda\otimes\mathbf{L}_{\tilde N}\mu.\end{aligned} \tag{7.3}$$

If an e-tensor field is an intersection of a spacetime tensor field $\bar\lambda(X)$ by the hypersurface,

$$\lambda(x)[e] \equiv \bar\lambda(e(x)), \tag{7.4}$$

its Lie derivative $\mathbf{L}_{\tilde N}$ vanishes,

$$\mathbf{L}_{\tilde N}\lambda = 0. \tag{7.5}$$

In particular, the Lie derivative of the metric field $\underset{\sim}{g}$ vanishes,

$$\mathbf{L}_{\tilde N}\underset{\sim}{g}(e(x)) = 0. \tag{7.6}$$

Further, the Lie derivative $\mathbf{L}_{\tilde N}$ annihilates the e-tensor fields $\bar n$ and $\underset{\sim}{\bar e}$ which define the normal e-basis. Namely, because $\delta_{\gamma x'}e_a^\alpha(x) = \delta^\alpha_\gamma\delta_{,a}(x,x')$,

$$\mathbf{L}_{\tilde N}e_a^\alpha(x) = N^c e^\alpha_{a,c} - (N^c e_c^\gamma)^{x'}\delta_{\gamma x'}e_a^\alpha + e_c^\alpha N^c{}_{,a} = 0.$$

The conclusion that $\mathbf{L}_{\tilde N}n_\alpha(x) = 0$ is reached by applying $\mathbf{L}_{\tilde N}$ to the definition equations of n_α,

$$g^{\alpha\beta}n_\alpha n_\beta = \epsilon, \quad e_a^\alpha n_\alpha = 0.$$

Because $\mathbf{L}_{\tilde N}g_{\alpha\beta} = \mathbf{L}_{\tilde N}e_a^\alpha = 0$, the $\mathbf{L}_{\tilde N}$ derivative of $g_{ab} = g_{\alpha\beta}e_a^\alpha e_b^\beta$ also vanishes. In summary, the $\mathbf{L}_{\tilde N}$ derivative of $\underset{\sim}{g}$, g, $\underset{\sim}{\bar e}$ and $\bar n$, considered as e-fields, vanishes,

$$\mathbf{L}_{\tilde N}\underset{\sim}{g} = \mathbf{L}_{\tilde N}g = \mathbf{L}_{\tilde N}\underset{\sim}{\bar e} = \mathbf{L}_{\tilde N}\bar n = 0. \tag{7.7}$$

The operation of the Lie derivative $\mathbf{L}_{\tilde N}$ thus commutes with all the natural operations on e-tensor fields, like raising and lowering of indices, contractions, projections of spacetime indices into $\perp$ and $\parallel$ directions to the hypersurface, and the lifting of a space tensor index into a spacetime tensor index.

Two embeddings, e and $\underset{\circ}{e}$, which differ only by a space diffeomorphism, $e = \underset{\circ}{e}\circ\varphi$, define the same hypersurface. If an e-tensor field is such that the expression in $\{\}$ in Eq. (7.1) vanishes for any φ, we can interpret it as a tensor field defined on hyperspace rather than on the space of embeddings. We shall call the e-tensors of such type *hypertensors*. The differential condition for λ to be a hypertensor obviously is that

$$\mathbf{L}_N\lambda = 0, \quad \forall\tilde N. \tag{7.8}$$

An intersection of a spacetime tensor field is a hypertensor field. The fields $\bar n$ and $\underset{\sim}{\bar e}$ are also hypertensor fields. The projections of a spacetime tensor field intersected by a hypersurface are therefore hypertensor fields. The sum of two hypertensor fields of the same rank or the direct product of two hypertensor fields is again a hypertensor field. Lowering or raising an arbitrary index of a hypertensor field, and projecting or lifting its indices from spacetime to space or vice versa, leads again to a hypertensor field.

Along with hypertensors, we can study the $/\!/$-tensors. A functional $f[e]$ is an $/\!/$-scalar if its value depends only on the hypersurface h, not on its particular representation by e,

$$f[e\circ\phi] = f[e]. \tag{7.9}$$

The differential version of this condition is

$$\delta_{ax}f[e] = 0. \tag{7.10}$$

An $\mathcal{E}$-vector $\mathbf{M}$ is characterized by the component e-vector $\bar M(x)$. We say that an $\mathcal{E}$-vector field $\mathbf{M}[e]$ is

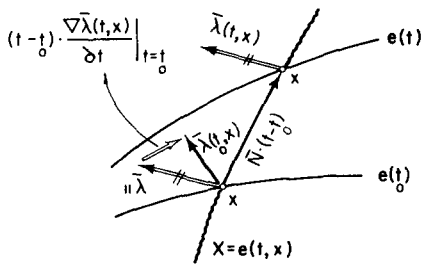

FIG. 3. Induced covariant derivative in $\mathcal{E}$. The covariant $\mathcal{E}$-derivative $\nabla_{\mathbf{N}}\boldsymbol{\lambda}$ of an e-vector field $\boldsymbol{\lambda}$ along the deformation $\mathcal{E}$-vector $\mathbf{N}$ is defined by means of the spacetime covariant derivative $\nabla/\partial t$ of the component field $\bar\lambda(t,x)$ along the spacetime curve $X=e(t,x)$, x fixed.

an H-vector field if the component e-vector field $\bar M(x)[e]$ is a hypervector field. From the definitions of the [ ] and the $\mathbf{L}_{\tilde N}$ operations, we get the identity

$$N^{ax'}[\delta_{ax'},\mathbf{M}]=-\mathbf{L}_{\tilde N}\mathbf{M}. \tag{7.11}$$

Therefore, $\mathbf{M}$ is an H-vector field iff

$$[\delta_{ax'},\mathbf{M}]=0. \tag{7.12}$$

The last equation is the differential version of the condition

$$\mathbf{M}[e\circ\varphi]f[e]=\mathbf{M}[e]f[e],\quad \forall f. \tag{7.13}$$

Indeed, writing Eq. (7.13) in the form

$$\frac{1}{t}(\mathbf{M}[e]\,f[e]|_{e\circ\varphi_t}-\mathbf{M}[e]\,f[e])$$

$$-\mathbf{M}[e\circ\varphi_t]\frac{1}{t}(f[e\circ\varphi_t]-f[e])=0 \tag{7.14}$$

and passing to the limit $t\to 0$, we get exactly

$$N^{ax'}\delta_{ax'}(\mathbf{M}f)-\mathbf{M}(N^{ax'}\delta_{ax'}f)=N^{ax'}[\delta_{ax'},\mathbf{M}]f=0.$$

The space vector $N^a(x)$ is the same for all embeddings, being defined solely by the diffeomorphisms φ_t on m.

Note that if $\mathbf{M}$ and $\mathbf{N}$ are H-vector fields, then $[\mathbf{M},\mathbf{N}]$ is also an H-vector field, due to the Jacobi identity for the Lie bracket [ ].

One can extend the definition of H-vectors to H-covectors and then to arbitrary H-tensors, but we shall have little opportunity to work with H-tensors in the following.

8. INDUCED COVARIANT DIFFERENTIATION IN HYPERSPACE

Covariant differentiation of spacetime vectors directly induces a covariant differentiation ∇ in hyperspace. We define first the covariant derivative $\nabla_{\mathbf{N}}\mathbf{M}$ of an $\mathcal{E}$-vector field $\mathbf{M}$ along an $\mathcal{E}$-vector $\mathbf{N}$. The "local" character of the operation $\nabla_{\mathbf{N}}$ allows us then to extend if from $\mathcal{E}$-vector fields to e-tensor fields. Finally, we show that if $\mathbf{N}$ is an H-vector and $\mathbf{M}$ a hyperfield, $\nabla_{\mathbf{N}}\mathbf{M}$ is again a hyperfield.

Let $\mathbf{M}$ be an $\mathcal{E}$-vector field defined along a curve $e=e(t)$ and $\mathbf{N}$ the tangent $\mathcal{E}$-vector to this curve at the point $e(t_0)$. In a coordinate $\mathcal{E}$-basis, the $\mathcal{E}$-vectors $\mathbf{M}$ and $\mathbf{N}$ are characterized by the spacetime vector fields, $\bar M$ and $\bar N$, defined along the embedding:

$$\mathbf{M}[e(t)]\longrightarrow\bar M[e(t,x)],$$

$$\mathbf{N}[e(t_0)]\longrightarrow\bar N[e(t_0,x)]=\frac{\partial e(t,x)}{\partial t}\Big|_{t=t_0}.$$

Keeping $x\in m$ fixed, we can form the spacetime covariant derivative $\nabla\bar M/\partial t\,|_t$ of $\bar M$ along the curve $X=e(t,x)$ at the point $X=e(t_0,x)$. The spacetime vector field $[\nabla\bar M/\partial t]_{t_0}$ generated in this way along the embedding $X=e(t_0,x)$ is then taken, by definition, to be the component of the covariant derivative $\nabla_{\mathbf{N}}$ with respect to the coordinate $\mathcal{E}$-basis (Fig. 3):

$$\nabla_{\mathbf{N}}\mathbf{M}|_{e(t_0)}\longrightarrow\frac{\nabla\bar M}{\partial t}\Big|_{t_0}. \tag{8.1}$$

In a concrete coordinate $\mathcal{E}$-basis $\delta_{\alpha x}$, this prescription gives

$$[\nabla_{\mathbf{N}}\mathbf{M}]^{\alpha x}\equiv\nabla_{\mathbf{N}}M^{\alpha x}=\frac{\partial M^\alpha(t,x)}{\partial t}\Big|_{t=t_0}$$

$$+\Gamma^\alpha{}_{\beta\gamma}(e(t_0,x))\,M^\beta(t_0,x)\,N^\gamma(t_0,x), \tag{8.2}$$

where $\Gamma^\alpha{}_{\beta\gamma}(X)$ is the Riemannian connection generated by the spacetime metric $g_{\alpha\beta}(X)$. When $\mathbf{M}$ is defined not only along the curve $e=e(t)$, but in a region of $\mathcal{E}$ around the point $e(t_0)$, we have

$$M^\alpha(t,x)=M^\alpha[e]_{e(t,x)}, \tag{8.3}$$

and the formula (8.2) gives

$$[\nabla_{\mathbf{N}}\mathbf{M}]^{\alpha x}=N^{\gamma x'}(\delta_{\gamma x'}M^\alpha[e(x)]$$

$$+\Gamma^\alpha{}_{\beta\gamma}(e(x))\,M^\beta[e(x)]\,\delta_{xx'}). \tag{8.4}$$

The coefficient of $N^{\gamma x'}$ on the right-hand side of this equation may be interpreted as $\nabla_{\gamma x'}M^{\alpha x}$.

Equation (8.4) shows that the components of the affine $\mathcal{E}$-connection in the coordinate $\mathcal{E}$-basis are given by

$$\Gamma^{\alpha x}{}_{\beta x'\,\gamma x''}=\Gamma^\alpha{}_{\beta\gamma}(e(x))\,\delta(x,x')\,\delta(x,x''). \tag{8.5}$$

If $\mathbf{M}$ is an intersection of a spacetime vector field by the hypersurface,

$$\bar M[e(x)]=\bar M(e(x)),$$

the variational derivative $\delta_{\gamma x'}M^{\alpha x}$ is equal to $M^\alpha{}_{,\gamma}(e(x))\,\delta(x,x')$, and from Eq. (8.4) we get the formula

$$[\nabla_{\mathbf{N}}\mathbf{M}]^{\delta e(x)}=\nabla_{\bar N}\bar M|_{X=e(x)}. \tag{8.6}$$

The covariant $\mathcal{E}$-derivative then becomes an intersection of the spacetime covariant derivative by the embedding $X=e(x)$.

The covariant differentiation in $\mathcal{E}$ has all the standard properties of a covariant differentiation, namely

$$\nabla_{\mathbf{M}+\mathbf{N}}\mathbf{P}=\nabla_{\mathbf{M}}\mathbf{P}+\nabla_{\mathbf{N}}\mathbf{P}, \tag{8.7}$$

$$\nabla_{\mathbf{N}}(\mathbf{M}+\mathbf{P})=\nabla_{\mathbf{N}}\mathbf{M}+\nabla_{\mathbf{N}}\mathbf{P}, \tag{8.8}$$

$$\nabla_{f\mathbf{N}}\mathbf{M}=f\nabla_{\mathbf{N}}\mathbf{M}, \tag{8.9}$$

$$\nabla_{\mathbf{N}}f\mathbf{M}=(\mathbf{N}f)\,\mathbf{M}+f\nabla_{\mathbf{N}}\mathbf{M}, \tag{8.10}$$

$$\nabla_{\mathbf{N}}\mathbf{M}-\nabla_{\mathbf{M}}\mathbf{N}=[\mathbf{N},\mathbf{M}]. \tag{8.11}$$

Here, $\mathbf{M},\mathbf{N},\mathbf{P}$ are arbitrary $\mathcal{E}$-vector fields, and $f\in\mathcal{F}(\mathcal{E})$ is an arbitrary functional on $\mathcal{E}$.

There is no single regular metric $\mathbf{g}$ in $T(\mathcal{E})$ which would be covariantly conserved by ∇. However, we can define an infinity of degenerate metrics $\mathbf{g}_x$ in $T(\mathcal{E})$, one per each point $x\in m$ of the embedding, which are covariantly conserved by ∇,

$$\nabla_{\mathbf{N}}\mathbf{g}_x=0\quad\forall\mathbf{N}. \tag{8.12}$$

These metrics are given by the prescription

$$\mathbf{g}_x(\mathbf{M},\mathbf{N})=\underset{\sim}{g}(\bar{M},\bar{N})\big|_{X=e(x)}, \tag{8.13}$$

their (infinite) degeneracy being obvious from the fact that

$$\mathbf{g}_x(.\,,\mathbf{N})=0$$

for any $\mathcal{E}$-vector $\mathbf{N}$ such that $\bar{N}(x)=0$. With respect to a coordinate $\mathcal{E}$-cobasis $\mathbf{d}e^{\alpha x'}$, $\mathbf{g}_x$ has the components

$$g_{x\,\alpha x'\,\beta x''}=g_{\alpha\beta}(e(x))\,\delta(x,x')\,\delta(x,x''). \tag{8.14}$$

Conversely, the rules (8.7)–(8.12) determine the covariant differentiation in $\mathcal{E}$ uniquely, and we can recover from them the prescriptions (8.2)–(8.6) by standard methods.

The definition (8.1) of the covariant derivative in $\mathcal{E}$ may be extended to an arbitrary e-tensor field $\lambda\in\mathcal{T}^{R,r}_{S,s}(\mathcal{E})$ along $e(t)$. We put

$$\nabla_{\mathbf{N}}\lambda\big|_{e(\overset{t}{\circ})}=\frac{\nabla\tilde{\lambda}}{\partial t}\bigg|_{\overset{t}{\circ}}. \tag{8.15}$$

The covariant derivative $\nabla/\partial t$ applies only to the spacetime leg $\binom{R}{S}$ of λ, treating the space leg $\binom{r}{s}$ of λ as a scalar in $\mathcal{H}$. Because $\lambda(t,x)$ is a space tensor in $T_x{}^r_s(m)$ for every t, so is $[\nabla\tilde{\lambda}/\partial t]_{\overset{t}{\circ}}$. The operation $\nabla_{\mathbf{N}}$ thus maps a field of e-tensors along $e(t)$ into an e-tensor at $e(\overset{t}{\circ})$ of the same rank. If λ is defined in a region of $\mathcal{E}$ around $e(\overset{t}{\circ})$, the coordinate expression of $\nabla_{\mathbf{N}}\lambda$ is

$$\begin{aligned}[\nabla_{\mathbf{N}}\lambda]^{\alpha\cdots a\cdots}{}_{\beta\cdots b\cdots}(x)&\equiv\nabla_{\mathbf{N}}\lambda^{\alpha\cdots a\cdots}{}_{\beta\cdots b\cdots}(x)\\&=N^{\gamma x'}\delta_{\gamma x'}\lambda^{\alpha\cdots a\cdots}{}_{\beta\cdots b\cdots}(x)\\&\quad+\Gamma^{\alpha}{}_{\delta\gamma}(e(x))\,\lambda^{\delta\cdots a\cdots}{}_{\beta\cdots b\cdots}(x)\,N^{\gamma}(x)+\cdots\\&\quad-\Gamma^{\delta}{}_{\beta\gamma}(e(x))\,\lambda^{\alpha\cdots a\cdots}{}_{\delta\cdots b\cdots}(x)\,N^{\gamma}(x)-\cdots.\end{aligned} \tag{8.16}$$

The covariant derivative ∇ is a derivation on the algebra of e-tensors, because

$$\nabla_{\mathbf{N}}(\lambda\otimes\mu)=\nabla_{\mathbf{N}}\lambda\otimes\mu+\lambda\otimes\nabla_{\mathbf{N}}\mu. \tag{8.17}$$

Note the local character of the direct product $\otimes$; to get $\lambda\otimes\mu$, the e-tensors are multiplied pointwise for $x\in m$. On the other hand, the direct product of two $\mathcal{E}$-tensors is nonlocal. The covariant derivative ∇ acts as a derivation with respect to the local direct product (8.17) because the affine connection (8.5) is local in the pair x,x', containing the δ function $\delta(x,x')$.

The covariant derivative ∇ commutes with the raising and lowering of the spacetime indices and the spacetime contractions. It does not commute, however, with the other natural algebraic operations on e-tensors: The raising and lowering of the space indices, space contractions, the projections into the normal and tangential directions to the hypersurface, and the lifting of space tensors into spacetime tensors. In Sec. 11, we shall define a new (nonsymmetrical) covariant derivative $\overset{*}{\nabla}$ which will commute with all these operations.

To prove that the covariant derivative $\nabla_{\mathbf{N}}$ may be interpreted as a covariant derivative in hyperspace, we turn to its commutation relation with the Lie derivative operation $\mathbf{L}_{\tilde{M}}$. Applied to an arbitrary e-tensor field,

$$\mathbf{L}_{\tilde{M}}\nabla_{\mathbf{N}}-\nabla_{\mathbf{N}}\mathbf{L}_{\tilde{M}}=\nabla_{\mathbf{L}_{\tilde{M}}\mathbf{N}}. \tag{8.18}$$

Equation (8.18) may be checked directly from the definitions (8.16) of $\nabla_{\mathbf{N}}$ and (7.2) of $\mathbf{L}_{\tilde{M}}$, due regard being given to the "intersection" nature (8.5) of the affine connection. From Eq. (8.18) we immediately see that if $\mathbf{N}$ is an $\mathcal{H}$-vector and λ a hypertensor field, the covariant derivative $\nabla_{\mathbf{N}}\lambda$ is again a hypertensor field. We can thus regard $\nabla_{\mathbf{N}}$ as a covariant derivative in hyperspace.

9. DEFORMATIONS OF THE NORMAL HYPERBASIS

Notable examples of hypervector fields which are not intersections of spacetime vector fields are the normal and tangential hypervector fields $\mathbf{n}\multimap\bar{n}(x)$ and $\mathbf{e}\multimap\underset{\sim}{\bar{e}}(x)$. We shall study now how these fields change when we pass from one hypersurface to another along a curve in hyperspace. The natural measure of this change is the covariant hyperderivative $\nabla_{\mathbf{N}}$ of $\mathbf{n}$ and $\mathbf{e}$ along the deformation hypervector

$$\mathbf{N}\big|_{e(\overset{t}{\circ})}\multimap\bar{N}=\frac{\partial e(x,t)}{\partial t}\bigg|_{t=\overset{t}{\circ}}. \tag{9.1}$$

The hypervectors $\nabla_{\mathbf{N}}\mathbf{n}$ and $\nabla_{\mathbf{N}}\mathbf{e}$ at the hypersurface $\{e(t)\}$ are characterized by the components $[\nabla_{\mathbf{N}}\mathbf{n}]^{\delta e(x)}$ and $[\nabla_{\mathbf{N}}\mathbf{e}]^{\delta e(x)}$. Here, $[\nabla_{\mathbf{N}}\mathbf{n}]^{\delta e(x)}$ is the field of spacetime vectors, and $[\nabla_{\mathbf{N}}\mathbf{e}]^{\delta e(x)}$ is the field of spacetime vectors—space covectors along the hypersurface $\{e(t)\}$. The geometrical meaning of these two fields is illustrated in Fig. 4.

The fields $[\nabla_{\mathbf{N}}\mathbf{n}]^{\delta e(x)}$ and $[\nabla_{\mathbf{N}}\mathbf{e}]^{\delta e(x)}$ depend only on the deformation field $\bar{N}$, not on the details of the further

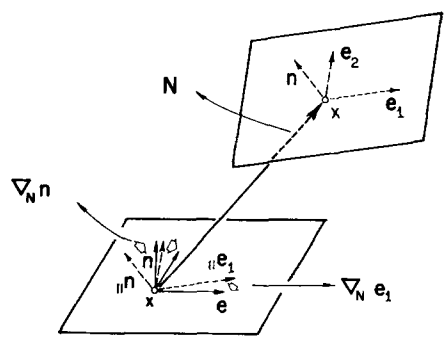

FIG. 4. Deformation of the normal hyperbasis. The normal hyperbasis $\{\mathbf{n},\mathbf{e}\}$ is parallel transported from the deformed hypersurface to the original hypersurface and compared with the original hyperbasis.

course of the curve $\{e(t)\}$. However, they depend not only on the value of $\bar{N}$ at the point x, but also on the behavior of $\bar{N}$ in the neighborhood of x. This shows that $\nabla_{\mathbf{N}}\mathbf{n}$ and $\nabla_{\mathbf{N}}\mathbf{e}$ are covariant hyperderivatives of genuine hypervector fields and not simple intersections of spacetime covariant derivatives of some spacetime vector fields.

We shall determine first the field $[\nabla_{\mathbf{N}}\mathbf{e}]^{\alpha}_{a}(x) \equiv \nabla_{\mathbf{N}} e^{\alpha}_{a}(x)$. Because the spacetime covariant derivatives ∇ along the t-lines and along the hypersurface are interchangeable,

$$\nabla_{\mathbf{N}} e^{\alpha}_{a}(x) = \frac{\nabla e^{\alpha}_{a}(x,t)}{\partial t} = \frac{\nabla}{\partial x^{a}}\frac{\partial e^{\alpha}}{\partial t} = \frac{\nabla N^{\alpha}}{\partial x^{a}}, \tag{9.2}$$

we get $\nabla_{\mathbf{N}} e^{\alpha}_{a}$ by substituting the lapse-shift decomposition (4.14), (4.15) into Eq. (9.2);

$$\frac{\nabla N^{\alpha}}{\partial x^{a}} = N_{,a}n^{\alpha} + N\frac{\nabla n^{\alpha}}{\partial x^{a}} + N^{b}{}_{,a}e^{\alpha}_{b} + N^{b}\frac{\nabla e^{\alpha}_{b}}{\partial x^{a}}.$$

However, we already know how the basis $\{n^{\alpha}(x), e^{\alpha}_{b}(x)\}$ changes along the hypersurface from Eq. (3.24) and (3.27). We thus get

$$\nabla_{\mathbf{N}} e^{\alpha}_{a}(x) = (N_{,a} + \epsilon K_{ab}N^{b})n^{\alpha} + (-K^{b}_{a}N + N^{b}{}_{|a})e^{\alpha}_{b}. \tag{9.3}$$

The hyperderivative $[\nabla_{\mathbf{N}}\mathbf{n}]^{\delta e(x)}$ is then determined from the orthogonality relations

$$n_{\alpha}e^{\alpha}_{a} = 0, \quad g^{\alpha\beta}n_{\alpha}n_{\beta} = \epsilon. \tag{9.4}$$

Applying to them the covariant derivative $\nabla/\partial t$, we get

$$e^{\alpha}_{a}\frac{\nabla n_{\alpha}}{\partial t} = -n_{\alpha}\frac{\nabla e^{\alpha}_{a}}{\partial t}, \quad n^{\alpha}\frac{\nabla n_{\alpha}}{\partial t} = 0. \tag{9.5}$$

Therefore, immediately,

$$[\nabla_{\mathbf{N}}\mathbf{n}]^{\alpha x} = -(\epsilon N_{,a} + K_{ab}N^{b})e^{a\alpha}. \tag{9.6}$$

Equations (9.3) and (9.6) express the deformation of the hyperbasis $\{\mathbf{n}, \mathbf{e}\} \rightarrowtail \{\bar{n}(x), \underset{\sim}{\bar{e}}(x)\}$ under the deformation $\mathbf{N}$ of the hypersurface.

While the index α can be raised and lowered in Eqs. (9.3) and (9.6) behind the symbol $\nabla_{\mathbf{N}}$, this is not so with the space index a in Eq. (9.3). The space metric tensor g_{ab} changes under the deformation $\mathbf{N}$,

$$\nabla_{\mathbf{N}} g_{ab}(x) = \nabla_{\mathbf{N}}(g_{\alpha\beta}e^{\alpha}_{a}e^{\beta}_{b}) = g_{\alpha\beta}\nabla_{\mathbf{N}}(e^{\alpha}_{a}e^{\beta}_{b}) = 2e_{\alpha(b}\nabla_{\mathbf{N}}e^{\alpha}_{a)}.$$

Substituting here for $\nabla_{\mathbf{N}} e^{\alpha}_{a}$ from Eq. (9.3), we get an important formula

$$\nabla_{\mathbf{N}} g_{ab} = -2K_{ab}N + 2N_{(a|b)}. \tag{9.7}$$

Similarly,

$$\nabla_{\mathbf{N}} g^{ab} = -g^{ac}g^{bd}\nabla_{\mathbf{N}} g_{cd} = -2K^{ab}N - 2N^{(a|b)}. \tag{9.8}$$

Finally, we get the $\nabla_{\mathbf{N}}$ derivative of the Levi-Civita form $\eta_{abc} = g^{1/2}\delta_{abc}$ considered as a hypervector,

$$\nabla_{\mathbf{N}} g^{1/2} = \tfrac{1}{2}g^{1/2}g^{ab}\nabla_{\mathbf{N}} g_{ab} = (-KN + N^{a}{}_{|a})g^{1/2}, \qquad \nabla_{\mathbf{N}}\eta = (-KN + N^{a}{}_{|a})\eta. \tag{9.9}$$

From Eqs. (9.3) and (9.8), we can find the change of $e^{a}_{\alpha} \equiv g_{\alpha\beta}g^{ab}e^{\beta}_{b}$ under the deformation $\mathbf{N}$,

$$\nabla_{\mathbf{N}} e^{a}_{\alpha} = (N_{,}{}^{a} + \epsilon K^{a}_{b}N^{b} - N^{a}{}_{|b})e^{b}_{\alpha}. \tag{9.10}$$

Note once more an important difference between Eqs. (9.3), (9.6), (9.10) and those equations one would expect to obtain if $\nabla_{\mathbf{N}}$ was a spacetime covariant derivative $\nabla_{\bar{N}}$ of a tensor field which follows. The right-hand sides of the mentioned equations depend on the derivatives of $\bar{N}(x)$, as exemplified by the presence of terms $N_{,a}$ or $N^{a}{}_{|b}$. This is because the vector fields $\bar{n}(x)$ and $\underset{\sim}{\bar{e}}(x)$ depend not only on the position in spacetime, but also on the hypersurface itself.

It is useful to split the basic equations (9.3), (9.6), (9.7) of this section into deformations which are normal and tangential to the hypersurface. Denoting by the symbols ∇_{N} and $\nabla_{\tilde{N}}$ the covariant hyperderivatives along the $/\!/$-vectors $\mathbf{N} = (Nn)^{\alpha x}\delta_{\alpha x} = \epsilon N^{x}\delta_{\perp x}$ and $\mathbf{N} = (N^{a}e^{\alpha}_{a})^{x}\delta_{\alpha x} = N^{ax}\delta_{ax}$, respectively, we get

$$\nabla_{N} n^{\alpha} = -\epsilon e^{a\alpha}N_{,a}, \quad \nabla_{N} e^{\alpha}_{a} = n^{\alpha}N_{,a} - K^{b}_{a}e^{\alpha}_{b}N \tag{9.11}$$

for the normal deformation of the hyperbasis $\{\mathbf{n}, \mathbf{e}\}$, and

$$\nabla_{\tilde{N}} n^{\alpha} = -K_{ab}e^{a\alpha}N^{b}, \qquad \nabla_{\tilde{N}} e^{\alpha}_{a} = \epsilon K_{ab}n^{\alpha}N^{b} + N^{b}{}_{|a}e^{\alpha}_{b} \tag{9.12}$$

for its tangential deformation.

The difference between Eqs. (9.12) and the corresponding Eqs. (3.24), (3.27) is easy to understand. The expressions $n_{\alpha;b}$ and $e^{\alpha}_{a;b}$ in Eqs. (3.24), (3.27) are covariant derivatives, along the tangent vectors $\bar{e}_{a}$, of the spacetime vector fields $n^{\alpha}(x)$ and $e^{\alpha}_{a}(x)$ defined along a prescribed embedding. The embedding $e: x \rightarrowtail X$ remains fixed. On the other hand, the operation $\nabla_{\tilde{N}}$ requires that the embedding itself be changed by $\bar{N} \equiv \partial e(t,x)/\partial t = \langle \underset{\sim}{\bar{e}}, \tilde{N}\rangle$, though the overall position of the hypersurface in spacetime remains fixed. Because the fields on which $\nabla_{\tilde{N}}$ operates are hyperfields, the operation $\nabla_{\tilde{N}}$ differs from the $\nabla_{\tilde{N}} \equiv N^{b}\nabla_{b}$ operation by a Lie derivative term. For a space scalar, like $\bar{n}$, we simply get $\nabla_{\tilde{N}}\mathbf{n}$. For a space covector, like $\underset{\sim}{\bar{e}}$, we have a more complicated relation

$$\nabla_{\tilde{N}} e^{\alpha}_{a} = \nabla_{\tilde{N}} e^{\alpha}_{a} - e^{\alpha}_{b}\nabla_{a}N^{b}. \tag{9.13}$$

Returning to Eq. (9.7), we can write down the normal and tangential deformations of the metric field $\underset{\sim}{g}$,

$$\nabla_{N}\underset{\sim}{g} = -2\underset{\sim}{K}N, \tag{9.14}$$

$$\nabla_{\tilde{N}}\underset{\sim}{g} = L_{\tilde{N}}\underset{\sim}{g}. \tag{9.15}$$

The first equation provides an alternative definition of the extrinsic curvature $\underset{\sim}{K}$, in addition to those given by formulas (3.23) and (3.25). The second equation equals the tangential deformation of $\underset{\sim}{g}$ to the Lie derivative of $\underset{\sim}{g}$ along the vector $\tilde{N}$. We can anticipate that the same relation between $\nabla_{\tilde{N}}$ and $L_{\tilde{N}}$ holds for an arbitrary projection of an arbitrary spacetime tensor.

From Eqs. (9.11) and (9.12), we can read the covariant hyperderivatives of the hyperbasis $\{\mathbf{n}, \mathbf{e}\}$ along the normal e-basis vectors $\delta_{\perp x}$ and δ_{bx}. We get

$$\nabla_{\perp x'} n^{\alpha}(x) = -e^{a\alpha}(x)\delta_{,a}(x,x'), \qquad \nabla_{\perp x'} e^{\alpha}_{a}(x) = -K^{b}_{a}e^{\alpha}_{b}\delta(x,x') + \epsilon n^{\alpha}\delta_{,a}(x,x'), \tag{9.16}$$

and

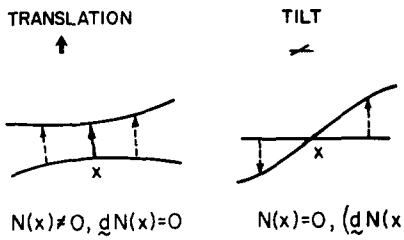

FIG. 5. Tilts and translations. A hypersurface translation is compared with a hypersurface tilt at the point $X=e(x)$.

$$\nabla_{bx'} n^\alpha(x) = -K_{ab}\, e^{a\alpha}\delta(x,x'),$$
$$\nabla_{bx'} e^\alpha_a(x) = \epsilon K_{ab} n^\alpha \delta(x,x') + e^\alpha_b(x)\,\delta_{,a}(x,x') + e^\alpha_c \gamma^c{}_{ba}\,\delta(x,x'). \qquad (9.17)$$

10. TILTS AND TRANSLATIONS OF HYPERSURFACES

The normal change ∇_N of the hyperbasis $\{n^\alpha, e^\alpha_a\}$ consists of two parts, $\nabla_\perp$ and $\nabla_\shortparallel$:

$$\nabla_\perp n^\alpha = -\epsilon e^{a\alpha} N_{,a}, \quad \nabla_\perp e^\alpha_a = n^\alpha N_{,a}, \qquad (10.1)$$

$$\nabla_\shortparallel n^\alpha = 0, \quad \nabla_\shortparallel e^\alpha_a = -K^b_a e^\alpha_b N. \qquad (10.2)$$

The change $\nabla_\shortparallel$ is local in the lapse function, depending only on the value of N at the point $x\in m$ in question, while the change $\nabla_\perp$ is nonlocal in the lapse function, depending on the gradient $\underset{\sim}{d}N$ of N. The deformation $\bar{N}=N\bar{n}$ will be called a *hypersurface tilt* at $x\in m$, if $N(x)=0$; it will be called a *hypersurface translation* at $x\in m$, if $\underset{\sim}{d}N|_x=0$ (Fig. 5). The tilts leave the spacetime point $X=e(x)$ fixed; the translations displace it to a new position.

The translation induces an affine transformation of the three tangent vectors $\bar{e}_a\in T_{X=e(x)}(\mathcal{M})$, leaving the normal vector $\bar{n}\in T_{X=e(x)}(\mathcal{M})$ fixed. Geometrically, the translation displaces the tangent plane of the hypersurface parallel to itself from the point $X=e(x)$ to the point $X+\bar{n}N\Delta t$; then, a new coordinate basis $\bar{e}_a$ is chosen at the new tangent plane, due to the new identification of points on the translated hypersurface induced by the new embedding $X+\bar{n}N\Delta t$ (Fig. 6).

The hypersurface tilt represents a Lorentz transformation (pure rotation) of the tangent space $T_{X=e(x)}(\mathcal{M})$ at a fixed spacetime point $X=e(x)$, because the lengths of the basis vectors and the angles between them are preserved by the tilt (Fig. 7). It is obvious from the construction of the deformed vectors that $\bar{n}$ remains a unit vector and $\bar{g}\rfloor\underline{n}=0$. The magnitudes and angles of the tangent vectors $\bar{e}_a$ and also preserved by the tilt, because

$$\nabla_\perp \underline{g}(\bar{e}_a,\bar{e}_b) = \nabla_\perp g_{ab} = 0$$

according to Eq. (8.7).

The behavior of the $\perp$ and $\shortparallel$ projections of a tensor field under a hypersurface tilt is completely predictable from the tensor character of that field alone. For this reason, the hypersurface tilts play the basic role in tensor kinematics. On the other hand, before telling what a tensor field does under a hypersurface translation, we must know the Lagrangian which determines the field dynamics. These relations are spelled out in the subsequent papers.

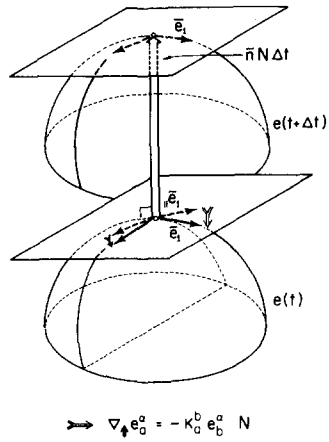

FIG. 6. Affine transformation of e^α_a under a hypersurface translation. A hypersurface translation displaces the tangent plane parallel to itself from the point $X=e(x)$ to the point $X+\bar{n}N\Delta t$. The coordinate basis $\bar{e}_a$ is parallel transported from the translated embedding to the original embedding and compared with the original basis.

11. NATURAL COVARIANT DIFFERENTIATION IN HYPERSPACE

The hyperspace covariant derivative ∇ which we have introduced in Sec. 8 was essentially the spacetime covariant derivative ∇ applied as directly as possible to hypertensors. The close connection between the derivatives ∇ and ∇ is reflected in the relation (8.5) between the affine connection $\Gamma^{\alpha x}{}_{\beta x'\,\gamma x''}$ and $\Gamma^\alpha{}_{\beta\gamma}$, and in the equality (8.6) of the two covariant derivatives ∇ and ∇ when applied to a spacetime vector field. When operating on hypertensors, however, the covariant derivative ∇ has the disadvantage that it does not commute with the raising and lowering of the space indices, with the projections of spacetime tensors into normal and tangential directions to the hypersurface, and with the lifting of a space tensor into a spacetime tensor. In this section, we shall introduce another covariant derivative $\overset{*}{\nabla}$ in the fiber of hypertensors which has all these desired properties. The new covariant derivative $\overset{*}{\nabla}$ will thus leave the normal e-basis field $\{n^\alpha, e^\alpha_a\}$ parallel propagated,

$$\overset{*}{\nabla}_N n^\alpha = 0, \quad \overset{*}{\nabla}_N e^\alpha_a = 0, \qquad (11.1)$$

and leave the metric field g_{ab} covariantly constant,

$$\overset{*}{\nabla}_N g_{ab} = 0. \qquad (11.2)$$

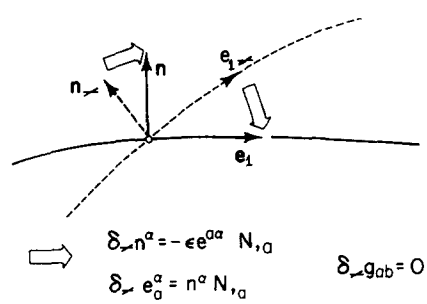

FIG. 7. Hypersurface tilt as a Lorentz transformation. Under a hypersurface tilt, the vectors $\{\bar{n},\bar{e}_a\}$ are rotated, their lengths and angles being preserved.

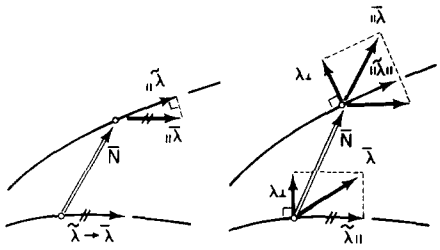

FIG. 8. Natural covariant derivative in $\mathcal{E}$. (a) A space e-vector $\boldsymbol{\lambda} \longleftrightarrow \tilde{\lambda}(x)$ is parallel transported by lifting it into a spacetime vector field $\bar{\lambda}(x)$ along the embedding, parallel transporting each spacetime vector $\lambda(x)$ along the deformation $\mathcal{E}$-vector $\mathbf{N} \longleftrightarrow \bar{N}(x)$, and projecting the parallel transported vector $_{\|}\bar{\lambda}$ back into the deformed embedding. (b) A spacetime e-vector $\lambda \longleftrightarrow \lambda(x)$ is decomposed into the tangential $\lambda_{\|}$ and normal $\lambda_{\perp}$ components. The tangential component is parallel transported as a space vector. The normal component is parallel transported by plotting $\lambda_{\perp}(x)$ along the normal to the deformed embedding. The parallel transported e-vector $_{\|}\boldsymbol{\lambda} \longleftrightarrow {}_{\|}\bar{\lambda}$ is the vector sum of $\langle \bar{e}, \lambda_{\|} \rangle$ and $\lambda_{\perp}\bar{n}$ at the deformed embedding.

As a consequence, it will also leave the metric field $g_{\alpha\beta}$ covariantly constant,

$$\overset{*}{\nabla}_{\mathbf{N}} g_{\alpha\beta} = 0. \tag{11.3}$$

The price we pay for the properties (11.1) and (11.2) is that the covariant derivative $\overset{*}{\nabla}$ is not a symmetrical covariant derivative, but has torsion. By and large, however, the price is worth the advantage which we gain when operating on e-tensor fields.

Unlike the symmetrical covariant derivative, the new covariant derivative does not treat the space legs of e-tensors as scalars, but operates on space indices as well as on the spacetime indices. In fact, the best starting point is to define first the covariant derivative $\overset{*}{\nabla}$ of an e-vector $\lambda^a(x)$.

We work in coordinate bases $\bar{\partial}_\alpha$, $\tilde{\partial}_a$, and $\delta_{\alpha x}$. For $\overset{*}{\nabla}_{\mathbf{N}}$ to be a local differentiation on the algebra of e-tensors, we require that the affine connections $\overset{*}{\Gamma}^{\alpha x}{}_{\beta x'' \gamma x'}$ and $\overset{*}{\gamma}{}^{ax}{}_{bx'' \gamma x'}$ act locally on the spacetime and space legs of e-tensors, respectively,

$$\overset{*}{\Gamma}^{\alpha x}{}_{\beta x'' \gamma x'} = \overset{*}{\Gamma}^{\alpha}{}_{\beta}(x)_{\gamma x'}\, \delta(x, x''),$$
$$\overset{*}{\gamma}{}^{ax}{}_{bx'' \gamma x'} = \overset{*}{\gamma}{}^{a}{}_{b}(x)_{\gamma x'}\, \delta(x, x'').$$

In other words,

$$\begin{aligned} \overset{*}{\nabla}_{\mathbf{N}} \lambda^\alpha(x) &= \delta_{\mathbf{N}} \lambda^\alpha(x) + \overset{*}{\Gamma}^{\alpha}{}_{\beta}(x)_{\gamma x'}\, \lambda^\beta(x) N^{\gamma x'}, \\ \overset{*}{\nabla}_{\mathbf{N}} \lambda^a(x) &= \delta_{\mathbf{N}} \lambda^a(x) + \overset{*}{\gamma}{}^{a}{}_{b}(x)_{\gamma x'}\, \lambda^b(x) N^{\gamma x'}. \end{aligned} \tag{11.4}$$

We define the parallel transport of an e-vector $\lambda^a(x)$ by lifting it into a spacetime vector field $\lambda^\alpha(x)$ along e, parallel propagating each vector $\lambda^\alpha(x)$ in the direction $\bar{N}(x) \longleftarrow \mathbf{N}$, and projecting the parallel propagated vectors back into the new hypersurface (Fig. 8). The definition of the parallel transport of $\lambda^a(x)$ thus closely follows the pattern used in the definition of the covariant derivative |, allowing, however, for the deformation of the hypersurface.

In terms of covariant derivatives, our definition of the parallel transport is expressed by the rule

$$\overset{*}{\nabla}_{\mathbf{N}} \lambda^a(x) \equiv e^a_\alpha \nabla_{\mathbf{N}} (\lambda^b e^\alpha_b). \tag{11.5}$$

Using Eq. (9.3), we get

$$\begin{aligned} \overset{*}{\nabla}_{\mathbf{N}} \lambda^a(x) &= e^a_\alpha e^\alpha_b \nabla_{\mathbf{N}} \lambda^b + e^a_\alpha \lambda^b \nabla_{\mathbf{N}} e^\alpha_b \\ &= \nabla_{\mathbf{N}} \lambda^a - K^a_b \lambda^b N + \lambda^b N^a{}_{|b} \\ &= \delta_{\mathbf{N}} \lambda^a + \int_{x' \in m} [-\epsilon K^a_b n_\gamma \delta(x, x') \\ &\quad + \gamma^a{}_{bc} e^c_\gamma \delta(x, x') + e^a_\gamma(x')\, \delta_{,b}(x, x')] \lambda^b(x) N^\gamma(x'). \end{aligned}$$

From here, we can identify the space leg of the affine connection,

$$\begin{aligned} \overset{*}{\gamma}{}^{a}{}_{b}(x)_{\gamma x'} \equiv \overset{*}{\gamma}{}^{a}{}_{b\gamma}(x, x') &= \gamma^a{}_{bc} e^c_\gamma(x, x') \\ &\quad - \epsilon K^a_b n_\gamma \delta(x, x') + e^a_\gamma(x')\, \delta_{,b}(x, x'). \end{aligned} \tag{11.6}$$

Using Eq. (9.7), one immediately verifies that the covariant derivative $\overset{*}{\nabla}_{\mathbf{N}}$ with the affine connection (11.6) leaves g_{ab} covariantly constant, Eq. (11.2).

The spacetime leg $\overset{*}{\Gamma}{}^{\alpha}{}_{\beta}(x)_{\gamma x'}$ of the affine connection is determined from the requirements (11.1). If we substitute into the equation

$$0 = \overset{*}{\nabla}_{\mathbf{N}} n^\alpha(x) = \nabla_{\mathbf{N}} n^\alpha(x) + [\overset{*}{\Gamma}{}^{\alpha}{}_{\beta}(x)_{\gamma x'} - \Gamma^\alpha{}_{\beta\gamma}\, \delta_{xx'}]\, n^\beta(x)\, N^{\gamma x'}, \tag{11.7}$$

the result (9.6), we can read off the $\overset{*}{\Gamma}{}^{\alpha}{}_{\perp}(x)_{\gamma x'}$ projection of the affine connection,

$$\begin{aligned} \overset{*}{\Gamma}{}^{\alpha}{}_{\perp}(x)_{\gamma x'} &= \Gamma^\alpha{}_{\perp\gamma}\, \delta(x, x') \\ &\quad + \epsilon K^a_c e^\alpha_a e^c_\gamma \delta(x, x') + \epsilon e^{a\alpha}(x)\, \delta_{,a}(x, x')\, n_\gamma(x'). \end{aligned} \tag{11.8}$$

Similarly, if we substitute into the equation

$$\begin{aligned} 0 = \overset{*}{\nabla}_{\mathbf{N}} e^\alpha_a &= \nabla_{\mathbf{N}} e^\alpha_a - \Gamma^\alpha{}_{\beta\gamma} e^\beta_a N^\gamma \\ &\quad + \overset{*}{\Gamma}{}^{\alpha}{}_{\beta}(x)_{\gamma x'}\, e^\beta_a(x)\, N^{\gamma x'} - \overset{*}{\gamma}{}^{b}{}_{a}(x)_{\gamma x'}\, e^\alpha_b(x)\, N^{\gamma x'} \end{aligned} \tag{11.9}$$

the already known connection $\overset{*}{\gamma}{}^{b}{}_{a}(x)_{\gamma x'}$ from Eq. (11.6) and the expression $\nabla_{\mathbf{N}}\, e^\alpha_a$ from Eq. (9.3), we can read off the $\overset{*}{\Gamma}{}^{\alpha}{}_{b}(x)_{\gamma x'}$ projection of the affine connection

$$\begin{aligned} \overset{*}{\Gamma}{}^{\alpha}{}_{b}(x)_{\gamma x'} &= \Gamma^\alpha{}_{b\gamma}\, \delta(x, x') - \epsilon K_{bc} n^\alpha e^c_\gamma \delta(x, x') \\ &\quad - \epsilon n^\alpha(x)\, n_\gamma(x')\, \delta_{,b}(x, x'). \end{aligned} \tag{11.10}$$

Putting the projections (11.8) and (11.10) together, we get the final result for the spacetime leg of the affine connection,

$$\begin{aligned} &\overset{*}{\Gamma}{}^{\alpha}{}_{\beta}(x)_{\gamma x'} \equiv \overset{*}{\Gamma}{}^{\alpha}{}_{\beta\gamma}(x, x') = \Gamma^\alpha{}_{\beta\gamma}(x, x') + \Lambda^\alpha{}_{\beta\gamma}(x, x'), \\ &\Lambda^\alpha{}_{\beta\gamma}(x, x') \equiv -\epsilon n^\alpha e^b_\beta e^c_\gamma K_{bc}\, \delta(x, x') \\ &\qquad + \epsilon e^\alpha_a n_\beta e^c_\gamma K^a_c\, \delta(x, x') \\ &\qquad + \epsilon e^{a\alpha} n_\beta n_\gamma(x')\, \delta_{,a}(x, x') \\ &\qquad - \epsilon n^\alpha e^a_\beta n_\gamma(x')\, \delta_{,a}(x, x'). \end{aligned} \tag{11.11}$$

The term $\Lambda^\alpha{}_{\beta\gamma}(x, x')$, being a difference of two affine connections, is an e-bitensor. It is conveniently catalogized by its projections (the last index being always projected into the e-basis at the point x'):

$$\Lambda^{\perp}{}_{b\perp} = -\delta_{,b}(x,x'),$$
$$\Lambda^{\perp}{}_{bc} = -\epsilon K_{bc}\,\delta(x,x'),$$
$$\Lambda^{a}{}_{\perp\perp} = \epsilon g^{ab}\,\delta_{,b}(x,x'), \tag{11.12}$$
$$\Lambda^{a}{}_{\perp c} = \epsilon K_{c}^{a}\,\delta(x,x').$$

All other projections of $\Lambda^{\alpha}{}_{\beta\gamma}$ are equal to zero.

Because $\mathbf{L}_{\tilde{M}}\,\Lambda^{\alpha}{}_{\beta\gamma} = 0$, $\Lambda^{\alpha}{}_{\beta\gamma}$ is actually a hyperbitensor. The covariant derivative $\overset{*}{\nabla}_{\mathbf{N}}$ thus satisfies the same commutation relation with $\mathbf{L}_{\tilde{M}}$ as $\nabla_{\mathbf{N}}$ does, namely

$$\mathbf{L}_{\tilde{M}}\overset{*}{\nabla}_{\mathbf{N}} - \overset{*}{\nabla}_{\mathbf{N}}\mathbf{L}_{\tilde{M}} = \overset{*}{\nabla}_{\mathbf{L}_{\tilde{M}}\mathbf{N}}. \tag{11.13}$$

Therefore, if $\mathbf{N}$ is an $\mathcal{H}$-vector, the covariant differentiation $\overset{*}{\nabla}_{\mathbf{N}}$ turns a hypervector field $\boldsymbol{\lambda}$ into a hypervector field $\overset{*}{\nabla}_{\mathbf{N}}\boldsymbol{\lambda}$.

The covariant hyperderivative $\overset{*}{\nabla}$ has again the standard properties (8.7)–(8.10) of covariant differentiation. Note that $f \in \mathcal{F}(\mathcal{E})$ must be a functional on $\mathcal{E}$, rather than a hyperscalar field, to get Eqs. (8.9), (8.10). Equation (8.11), however, must be modified. Due to the relation between the $\overset{*}{\nabla}_{\mathbf{N}}$ and $\nabla_{\mathbf{N}}$ derivatives, we get

$$\{\nabla_{\mathbf{N}}\mathbf{M} - \nabla_{\mathbf{M}}\mathbf{N} - [\mathbf{N},\mathbf{M}]\}^{\alpha x}$$
$$= \Lambda^{\alpha}{}_{\beta}(x)_{\gamma x'}\,(M^{\beta}(x)\,N^{\gamma x'} - N^{\beta}(x)\,M^{\gamma x'}). \tag{11.14}$$

The covariant derivative $\overset{*}{\nabla}_{\mathbf{N}}$ thus has torsion.

The basic trick of hypersurface dynamics is to project the covariant derivatives of spacetime tensors into normal and tangential components and express them by means of the hyperspace derivatives $\delta_{\perp x}$ and $\delta_{\parallel x}$ of these projections. For this purpose, any one of the covariant derivatives ∇ and $\overset{*}{\nabla}$ may be used in intermediary steps, though the ∇ derivative is easier to handle technically.

*Work supported in part by the National Science Foundation under Grant No. GP-43718X to the University of Utah.

†To the memory of my father.

[1]P.A.M. Dirac, Proc. R. Soc. London A **246**, 326, 333 (1958); Phys. Rev. **114**, 924 (1959); *Lectures on Quantum Mechanics* (Academic, New York, 1965).

[2]R. Arnowitt, S. Deser, and C.W. Misner, "The Dynamics of General Relativity," in *Gravitation: An Introduction to Current Research*, edited by L. Witten (Wiley, New York, 1962), and the original papers quoted there.

[3]For the geometrical approach to Hamiltonian geometrodynamics, see, e.g., A.E. Fisher and J.E. Marsden, J. Math. Phys. **13**, 546 (1972).

[4]B.S. DeWitt, Phys. Rev. **160**, 113 (1967); **162**, 1195, 1239 (1967).

[5]K. Kuchař, J. Math. Phys. **13**, 768 (1972).

[6]C. Teitelboim, Ann. Phys. 79, 542 (1973); K. Kuchař, "Canonical Quantization of Gravity," in *Relativity, Astrophysics and Cosmology*, edited by W. Israel (Reidel, Dordrecht, Holland, 1974).

[7]S.A. Hojman, K. Kuchař, and C. Teitelboim, Nature 245, 97 (1973); K. Kuchař, J. Math. Phys. **15**, 708 (1974).

Kinematics of tensor fields in hyperspace. II*

Karel Kuchař

Department of Physics, University of Utah, Salt Lake City, Utah 84112
(Received 14 July 1975)

Various kinematical relations, holding between hypersurface projections of spacetime tensor fields in an arbitrary Riemannian spacetime, are studied in terms of differential geometry in hyperspace. A criterion is given that a collection of hypertensor fields is generated by the projections of a single spacetime tensor field intersected by the embeddings. From here, it is shown that the super-Hamiltonian of an arbitrary tensor field splits into two parts, $H^{\phi}_{\uparrow}$ and $H^{\phi}_{\swarrow}$, $H^{\phi}_{\uparrow}$ being local in the field momenta and $H^{\phi}_{\swarrow}$ containing their first derivatives. The form of $H^{\phi}_{\swarrow}$ for an arbitrary tensor field is determined from the field behavior under hypersurface tilts. The kinematical equations for the intrinsic metric and the extrinsic curvature are written in a quasicanocical form, and their connection with the closing relations for the gravitational super-Hamiltonian is exhibited. The conservation laws of charge, energy and momentum, and the contracted Bianchi identities, are written as hypertensor equations.

1. INTRODUCTION

Hyperspace, which is an infinitely dimensional manifold of all spacelike hypersurfaces drawn in a Riemannian spacetime,[1] provides a natural way of looking at the dynamical evolution of tensor fields (including the metric field itself). One simply watches how the field changes when passing smoothly from one hypersurface to another along a curve in hyperspace. Ultimately, one aims at a dynamical scheme: specifying initially appropriate field variables on one hypersurface and determining them subsequently on any other hypersurface by means of the field equations. However, there are certain relations among the hypersurfaces, their intrinsic and extrinsic geometries, and the projections of various tensor fields, which hold in an arbitrary Riemannian spacetime containing arbitrarily distributed tensor fields. They hold irrespective of the dynamical laws which ultimately govern the fields in a dynamical theory and irrespective of the Einstein's law of gravitation which connects the spacetime curvature with the energy-momentum tensor of the fields. We call such relations *kinematical* (using this term well in conformity with classical mechanics) and devote the present paper to their study. Essentially, the kinematical relations hold because a single *spacetime* tensor field is projected into normal and tangential directions to an embedding, and these projections must behave in a definite way when the embedding is stretched, leaving the hypersurface fixed in spacetime, or when the hypersurface is tilted and bent along a fixed spacetime point. Despite their simple origin, the kinematical relations are essential for the correct visualization and interpretation of geometrodynamics, pure or driven by sources.

The formalism of the differential geometry in hyperspace, introduced in Ref. 1, is the main tool of our investigation. The notation is explained in Sec. 2 of that paper, and we quote its equations by prefixing the Roman numeral I before their section and equation numbers; (I·3·16), e.g., is Eq. (16) in Sec. 3 of Ref. 1.

The present paper is organized into ten sections. In Sec. 2, we use the connection between the spacetime covariant derivative ∇ and the covariant hyperderivative ∇, to express the projections of spacetime covariant derivatives $\nabla\bar{\lambda}$ of arbitrary spacetime tensor fields $\bar{\lambda}(X)$ in terms of the normal directional derivatives δ_N and the space covariant derivatives of the hyperfield projections $\lambda(x)[e]$. These projection formulas are the basic tool in casting the spacetime field equations into hyperfield equations. In Sec. 3, we ask the question when a collection of hypertensor fields may be thought about as generated by the projection of a single spacetime tensor field intersected by the embeddings. We find the answer to that question in the behavior of the collection under hypersurface tilts. In Sec. 4, we show that the super-Hamiltonian of any dynamical theory which describes the canonical evolution of a spacetime tensor field consists of two parts, $H^{\phi}_{\swarrow}$ and $H^{\phi}_{\uparrow}$. The first part, $H^{\phi}_{\swarrow}$, contains the derivatives of the field momenta, describes the behavior of the field under hypersurface tilts, and is completely determined by the kinematical considerations. The second part, $H^{\phi}_{\uparrow}$, is local in the field momenta, describes the truly dynamical evolution of the field under hypersurface translations, and is determined only by the specific Lagrangian which governs this evolution. The relation of the $H^{\phi}_{\swarrow}$ part to the spin energy–momentum tensor, and of the $H^{\phi}_{\uparrow}$ part to the canonical and symmetrical energy–momentum tensors, is discussed in the following paper. In Sec. 5, we show that the supermomentum of the field is likewise completely determined by kinematical considerations, from the behavior of the field projections under tangential deformations of the embedding. In Sec. 6, the Gauss–Codazzi equations and the normal deformation equation for the projections of the spacetime Riemann curvature tensor are written in the hyperspace language. The evolution equations for the intrinsic geometry $\underset{\sim}{g}$ and the extrinsic curvature $\underset{\sim}{K}$ of the hypersurface, which we obtain in this way, do not close, but their kinematical structure imposes a strong limitation on any possible geometrodynamics. In Sec. 7, we indicate how these equations lead to the closing relations between the super-Hamiltonians $H^{g}(x)$ of such geometrodynamics. This provides an additional insight into our earlier result[2] that the Einsteinian geometrodynamics is the unique canonical realization of these closing relations which uses the geometry $\underset{\sim}{g}$ as the sole configuration variable. In Sec. 8, the kinematical equations for $\underset{\sim}{g}$ and $\underset{\sim}{K}$ are cast into a quasicanonical form,

spoiled only by the presence of the G^{ab} projection of the Einstein's tensor, which plays the role of an external source. In Sec. 9, the conservation laws of charge, energy, and momentum are written as the evolution equations of these quantities in hyperspace. In Sec. 10, it is shown that the conservation laws of energy and momentum, when written as projected Bianchi identities for the Einstein's tensor $G^{\alpha\beta}$, take the form of the closing relations between the ADM super-Hamiltonians, and the ADM super-Hamiltonian and supermomentum.

Many results contained in this paper are the current results of Einsteinian geometrodynamics, incorporated into the new framework of hypertensor geometry. Thus, projecting the Riemann tensor is the standard device in the theory of embeddings, the kinematical equations for g and $\underset{\sim}{K}$ were cast into the quasicanonical form by Teitelboim,[3] the connection between the supermomenta and the Lie derivatives was known to Dirac[4] (who formulated it in the passive, rather than the active interpretation), and the connection between the closing relations for H^g and $H^g{}_a$, and the Bianchi identities was discussed by ADM.[5] On the other hand, the elucidation of the role which the hypersurface tilts play in hypersurface dynamics of tensor fields is, to our best knowledge, new. This applies in particular to the critierion when the collection of hypertensors represents a spacetime tensor field, and to the splitting of the super-Hamiltonian into the parts which are local and nonlocal in the field momenta, the nonlocal part describing the kinematical behavior of the tensor field under hypersurface tilts. The importance of $H^{\phi}_{\not\perp}$ was masked by the fact that $H^{\phi}_{\not\perp}$ vanishes for two of the most widely studied fields—the gravitational field and the scalar field—while for the third one, the electromagnetic field, it degenerates into the $\mathrm{div}\tilde{E}=0$ constraint.

2. PROJECTIONS OF SPACETIME COVARIANT DERIVATIVES

In hypersurface kinematics and dynamics of tensor fields, we need to express the projections of spacetime covariant derivatives of spacetime tensors in terms of hypersurface directional derivatives δ_N and the space covariant derivatives $|$ of the projections of these tensors. This is easily achieved by employing the properties of the induced covariant differentiation ∇ in hyperspace.

We take a spacetime covector field ϕ_α as an example. Studying first the normal covariant hyperderivative ∇_N of the projections $\phi_\perp$ and ϕ_a, we get

$$\begin{aligned}\nabla_N\phi_\perp &= \nabla_N(\epsilon n^\alpha\phi_\alpha)=\epsilon n^\alpha\nabla_N\phi_\alpha+\epsilon\phi_\alpha\nabla_N n^\alpha\\ &=\epsilon n^\alpha\phi_{\alpha;\beta}n^\beta N+\epsilon\phi_\alpha(-\epsilon e^{a\alpha}N_{,a})=\epsilon\phi_{\perp;\perp}N-\phi^a N_{,a}\end{aligned}\tag{2.1}$$

and

$$\begin{aligned}\nabla_N\phi_a &= \nabla_N(e^\alpha_a\phi_\alpha)=e^\alpha_a\nabla_N\phi_\alpha+\phi_\alpha\nabla_N e^\alpha_a\\ &=e^\alpha_a\phi_{\alpha;\beta}n^\beta N+\phi_\alpha(-K^b_a e^\alpha_b N+n^\alpha N_{,a})\\ &=\epsilon\phi_{a;\perp}N-K^b_a\phi_b N+\epsilon\phi_\perp N_{,a}\end{aligned}\tag{2.2}$$

by Eqs. (I.8.6) and (I.9.11) (see Ref. 1).

The tangential hyperderivatives $\nabla_{\tilde{N}}$ are equally easy to handle. We know from Sec. I.9 that the tangential hyperderivative of a hypertensor equals the Lie derivative $L_{\tilde{N}}$ of that hypertensor. Therefore, by Eq. (I.9.12),

$$\begin{aligned}N^b\phi_{\perp,b} &= L_{\tilde{N}}\phi_\perp=\nabla_{\tilde{N}}\phi_\perp=\nabla_{\tilde{N}}(\epsilon n^\alpha\phi_\alpha)\\ &=\epsilon n^\alpha\nabla_{\tilde{N}}\phi_\alpha+\epsilon\phi_\alpha\nabla_{\tilde{N}}n^\alpha\\ &=\epsilon n^\alpha\phi_{\alpha;\beta}e^\beta_b N^b+\epsilon\phi_\alpha(-K_{ab}e^{a\alpha}N^b)\\ &=(\phi_{\perp;b}-\epsilon K_{ab}\phi^a)N^b,\end{aligned}\tag{2.3}$$

and

$$\begin{aligned}&N^b\phi_{a|b}+\phi_b N^b{}_{|a}\\ &\quad=L_{\tilde{N}}\phi_a=\nabla_{\tilde{N}}\phi_a=\nabla_{\tilde{N}}(e^\alpha_a\phi_\alpha)\\ &\quad=e^\alpha_a\nabla_{\tilde{N}}\phi_\alpha+\phi_\alpha\nabla_{\tilde{N}}e^\alpha_a\\ &\quad=e^\alpha_a\phi_{\alpha;\beta}e^\beta_b N^b+\phi_\alpha(\epsilon K_{ab}n^\alpha N^b+N^b{}_{|a}e^\alpha_b)\\ &\quad=\phi_{a;b}N^b+\phi_\perp K_{ab}N^b+\phi_b N^b{}_{|a}.\end{aligned}\tag{2.4}$$

The shift vector N^b enters Eqs. (2.3) and (2.4) as an arbitrary multiplicative factor; therefore,

$$\phi_{\perp;b}=\phi_{\perp|b}+\epsilon K_{bc}\phi^c,\quad \phi_{a;b}=\phi_{a|b}-\phi_\perp K_{ab}.\tag{2.5}$$

In tangential projections, we need to know how the field changes along the hypersurface, not how it behaves when we pass to another hypersurface. On the other hand, the normal projections contain the directional derivatives δ_N,

$$\begin{aligned}\phi_{\perp;\perp}N&=\epsilon\,\delta_N\phi_\perp+\epsilon\phi^a N_{,a},\\ \phi_{a;\perp}N&=\epsilon\delta_N\phi_a+\epsilon K_{ab}\phi^b N-\phi_\perp N_{,a}.\end{aligned}\tag{2.6}$$

The projections of spacetime covariant derivatives of spacetime tensors $\phi_{\alpha\ldots\beta}$ are obtained when applying the rules (2.5) and (2.6) to each tensor index. As an example, the projection formulas for a covariant second-rank tensor $\phi_{\alpha\beta}$ give

$$\begin{aligned}\phi_{\perp\perp;c}&=\phi_{\perp\perp|c}+\epsilon K_{cd}\phi_\perp{}^d+\epsilon K_{cd}\phi^d{}_\perp,\\ \phi_{a\perp;c}&=\phi_{a\perp|c}-K_{ac}\phi_{\perp\perp}+\epsilon K_{cd}\phi_a{}^d,\\ \phi_{\perp b;c}&=\phi_{\perp b|c}-K_{bc}\phi_{\perp\perp}+\epsilon K_{cd}\phi^d{}_b,\\ \phi_{ab;c}&=\phi_{ab|c}-K_{ac}\phi_{\perp b}-K_{bc}\phi_{a\perp},\end{aligned}\tag{2.7}$$

and

$$\begin{aligned}\phi_{\perp\perp;\perp}N&=\epsilon\,\delta_N\phi_{\perp\perp}+\epsilon\phi^d{}_\perp N_{,d}+\epsilon\phi_\perp{}^d N_{,d},\\ \phi_{a\perp;\perp}N&=\epsilon\,\delta_N\phi_a+\epsilon K_{ad}\phi^d{}_\perp N+\epsilon\phi_a{}^d N_{,d}-\phi_{\perp\perp}N_{,a},\\ \phi_{\perp b;\perp}N&=\epsilon\,\delta_N\phi_{\perp b}+\epsilon K_{bd}\phi_\perp{}^d N+\epsilon\phi^d{}_b N_{,d}-\phi_{\perp\perp}N_{,b},\\ \phi_{ab;\perp}N&=\epsilon\,\delta_N\phi_{ab}+\epsilon K_{ad}\phi^d{}_b N+\epsilon K_{bd}\phi_a{}^d N\\ &\quad-\phi_{\perp b}N_{,a}-\phi_{a\perp}N_{,b}.\end{aligned}\tag{2.8}$$

3. HYPERSURFACE TILTS AND SPACETIME HYPERTENSORS

From Eqs. (2.6), we can read off the behavior of the $\phi_\perp$ and ϕ_a projections of a spacetime covector ϕ_α under the hypersurface tilts,

$$\delta_{\not\perp}\phi_\perp=-\phi^a N_{,a},\quad \delta_{\not\perp}\phi_a=\epsilon\phi_\perp N_{,a}.\tag{3.1}$$

Equations (3.1) tell us that the covector ϕ_α at the fixed spacetime point $\underset{o}{X}$ remains unchanged, but is projected

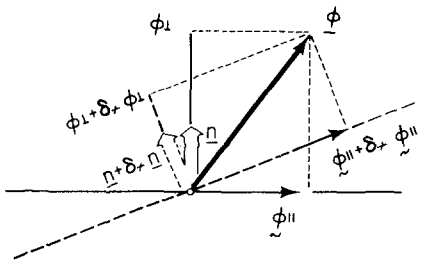

FIG. 1. Behavior of tensor projections under hypersurface tilts. The spacetime covector $\underset{\sim}{\phi}$ remains fixed, while its tangentail $\phi_{\|}$ and normal $\phi_{\perp}$ projections change under hypersurface tilts δ_{τ}.

into a tilted basis (I. 10. 1),

$$\delta_\tau n^\alpha = -\epsilon e^{a\alpha} N_{,a}, \quad \delta_\tau e^\alpha_a = n^\alpha N_{,a}, \tag{3.2}$$

(see Fig. 1).

Further, we know that the projections $\phi_\perp$ and ϕ_a of a spacetime covector are hypertensors (Sec. I. 7),

$$\mathbf{L}_{\tilde{N}}\ \phi_\perp = 0 = \mathbf{L}_{\tilde{N}}\ \phi_a. \tag{3.3}$$

Equations (3. 1) and (3. 3) thus hold for the projections of an arbitrary covector field $\phi_\alpha(X)$.

Let us ask now an inverse question, namely, when two e-tensor fields, $\phi_\perp(x)[e]$ and $\phi_a(x)[e]$, can be interpreted as the $\perp$ and $\|$ projections of a spacetime covector field $\phi_\alpha(X)$ intersected by the embeddings $X = e(x)$. The answer is again given by Eqs. (3. 1) and (3. 3), which are thus both necessary and sufficient conditions for an e-vector field

$$\phi_\alpha(x)[e] \equiv \phi_\perp(x)[e] n_\alpha + \phi_a(x)[e] e^a_\alpha \tag{3.4}$$

to be a spacetime hypervector field.

To prove that Eqs. (3. 1) and (3. 3) are sufficient conditions for the expression (3. 4) to be a spacetime hypervector, we must construct a spacetime covector field $\phi_\alpha(X)$, the intersection of which by an arbitrary embedding e gives the expression (3. 4). This we do point by point in $\mathcal{M}$, picking up an embedding $\underset{1}{e}$ which passes through the point $\underset{\circ}{X} \in \mathcal{M}$,

$$\underset{\circ}{X} = \underset{1}{e}(\underset{\circ}{x}), \quad \text{for a } \underset{\circ}{x} \in m, \tag{3.5}$$

defining

$$\phi_\alpha(\underset{\circ}{X}) \equiv \phi_\alpha(\underset{\circ}{x})[\underset{1}{e}], \tag{3.6}$$

and showing that $\phi_\alpha(\underset{\circ}{X})$ is the same for all embeddings $\underset{2}{e}$ passing through the same point $\underset{\circ}{X}$.

First, from Eq. (3. 3) it is obvious that $\phi_\alpha(\underset{\circ}{x})[\underset{1}{e}] = \phi_\alpha(\varphi(\underset{\circ}{x}))[\underset{1}{e} \circ \varphi^{-1}]$, i. e., $\phi_\alpha(\underset{\circ}{x})$ is the same for two embeddings, $\underset{1}{e}$ and $\underset{2}{e} = e \circ \varphi^{-1}$, which define the same hypersurface, at the given point $\underset{\circ}{X} = \underset{1}{e}(\underset{\circ}{x}) = \underset{2}{e}(\varphi(\underset{\circ}{x}))$ of that hypersurface (Sec. I. 7; Fig. 2).

Next, take an embedding $\underset{2}{e}$ which intersects the first embedding at $\underset{\circ}{X}$, but does not define the same hypersurface. Because we already know that the identification of space points along $\underset{2}{e}$ does not matter, we are free to choose $\underset{2}{e}$ so that the point $\underset{\circ}{X}$ corresponds to the same

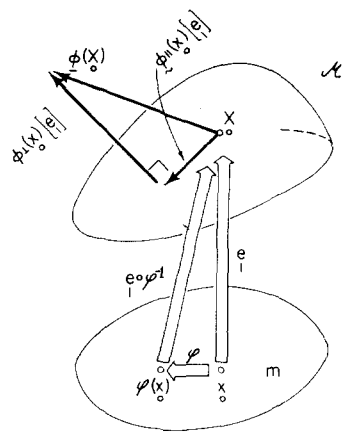

FIG. 2. Behavior of tensor projections under tangential deformations. The spacetime covector $\underset{\sim}{\phi}$ and its projections $\underset{\sim}{\phi}_{\|}$ and $\phi_\perp$ behave as hypertensors under a tangential deformation of the embedding.

point $\underset{\circ}{x} \in m$ according to the two embeddings $\underset{1}{e}$ and $\underset{2}{e}$ (Fig. 3),

$$\underset{1}{e}(\underset{\circ}{x}) = \underset{2}{e}(\underset{\circ}{x}). \tag{3.7}$$

We connect the embeddings $\underset{1}{e}$ and $\underset{2}{e}$ by a curve $e(t)$ in $\mathcal{E}$ such that

$$e(\underset{1}{t}) = \underset{1}{e}, \quad e(\underset{2}{t}) = \underset{2}{e}, \tag{3.8}$$

and

$$e(t, \underset{\circ}{x}) = \underset{\circ}{X} \quad \forall t \in (\underset{1}{t}, \underset{2}{t}). \tag{3.9}$$

Differentiating Eq. (3. 9) with respect to t, we see that the deformation vector field

$$\bar{N}(t, x) \equiv \frac{\partial e(t, x)}{\partial t} \longrightarrow \mathbf{N}(t) \tag{3.10}$$

vanishes at the point $\underset{\circ}{x} \in m$,

$$\bar{N}(t, \underset{\circ}{x}) = 0 \quad \forall t \in (\underset{1}{t}, \underset{2}{t}). \tag{3.11}$$

Decompose the vector field (3. 10) into the lapse function $N(t)$ and the shift vector $\tilde{N}(t)$. The rate of change of $\phi_\alpha(\underset{\circ}{x})[e(t)]$ along the curve $e(t)$ at the point $e = e(t)$ is given by

$$\frac{d}{dt}\phi_\alpha(\underset{\circ}{x})[e(t)] = \delta_N\, \phi_\alpha(\underset{\circ}{x})[e] + \delta_{\tilde{N}}\, \phi_\alpha(\underset{\circ}{x})[e]. \tag{3.12}$$

Equation (3. 3) ensures that $\delta_{\tilde{N}}\, \phi_\alpha(\underset{\circ}{x})[e] = 0$, because

$$0 = \mathbf{L}_{\tilde{N}}\, \phi_\alpha(\underset{\circ}{x})[e] = N^c(\underset{\circ}{x})\, \tilde{\partial}_c \lambda_\alpha(x)\big|_{\underset{\circ}{x}} - \delta_{\tilde{N}}\, \phi_\alpha(\underset{\circ}{x})[e] \tag{3.13}$$

according to Eq. (I. 7. 2), and $\tilde{N}(\underset{\circ}{x}) = 0$ by Eq. (3. 11). The change $\delta_N\, \phi_\alpha(\underset{\circ}{x})[e]$ is the tilt change δ_τ, because

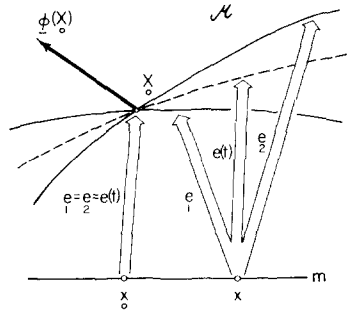

FIG. 3. Finite tilts. A finite tilt may be accomplished along a tilt curve $e(t)$ in $\mathcal{E}$.

$N(\overset{\circ}{x})=0$. Therefore,

$$\delta_N \phi_\alpha(\overset{\circ}{x})[e]=\delta_\tau(\phi_\perp(\overset{\circ}{x})[e]n_\alpha(\overset{\circ}{x})+\phi_a(\overset{\circ}{x})[e]e^a_\alpha(x))$$
$$=n_\alpha\delta_\tau\phi_\perp+\phi_\perp\delta_\tau n_\alpha+e^a_\alpha\delta_\tau\phi_a+\phi_a\delta_\tau e^a_\alpha=0, \tag{3.14}$$

by Eqs. (3.1) and (3.2), and because

$$\delta_\tau g_{\alpha\beta}=0=\delta_\tau g^{ab} \tag{3.15}$$

[cf. Eq. (I.9.8)]. This shows that $\phi_\alpha(\overset{\circ}{x})[e(t)]$ remains constant along the curve $e(t)$, so that $\phi_\alpha(\overset{\circ}{x})[\overset{2}{e}]=\phi_\alpha(\overset{\circ}{x})[\overset{1}{e}]\equiv\phi_\alpha(\overset{\circ}{X})$.

Our proof is thereby completed. It is straightforward to generalize the criterion (3.1), (3.3) so that it applies to an arbitrary spacetime tensor and its projections. For further use in geometrokinematics, let us write down the criterion (3.1) for a second-rank spacetime tensor $\phi_{\alpha\beta}$ [see Eq. (1.8)],

$$\begin{aligned}
&\delta_\tau\phi_{\perp\perp}=-(\phi^c{}_\perp+\phi_\perp{}^c)N_{,c},\\
&\delta_\tau\phi_{a\perp}=-\phi_a{}^cN_{,c}+\epsilon\phi_{\perp\perp}N_{,a},\\
&\delta_\tau\phi_{\perp b}=-\phi^c{}_bN_{,c}+\epsilon\phi_{\perp\perp}N_{,b},\\
&\delta_\tau\phi_{ab}=\epsilon\phi_{a\perp}N_{,b}+\epsilon\phi_{\perp b}N_{,a}.
\end{aligned} \tag{3.16}$$

4. CANONICAL REALIZATION OF HYPERSURFACE TILTS

The behavior of a given tensor field under hypersurface tilts is determined solely by the tensor character of the field in question. For example, the behavior of a vector field is given by Eqs. (3.1). On the other hand, the dynamical evolution of that field in any canonical theory is determined by the Poisson bracket of the field with a field "super-Hamiltonian"

$$H^\Phi(x)=H^\Phi(x)[\phi_\perp,\phi_a,\pi^\perp,\pi^a], \tag{4.1}$$

which is a functional of the field variables $\phi_\perp$, ϕ_a and their conjugate momenta $\pi^\perp$, π^a, according to the equation

$$\begin{aligned}
&\delta_N\phi_\perp(x)=[\phi_\perp(x),H^\Phi{}_{x'}]N^{x'},\\
&\delta_N\phi_a(x)=[\phi_a(x),H^\Phi{}_{x'}]N^{x'}.
\end{aligned} \tag{4.2}$$

At this stage, nothing further is known about the nature of the field momenta $\pi^\perp$, π^a, except that they are canonically conjugate to the field variables $\phi_\perp$, ϕ_a, satisfying thus the standard Poisson bracket relations

$$\begin{aligned}
&[\phi_\perp(x),\pi^\perp(x')]=\delta(x,x'),\\
&[\phi_a(x),\pi^b(x')]=\delta^b_a\,\delta(x,x'),
\end{aligned} \tag{4.3}$$

with all other fundamental Poisson brackets being equal to zero.

In particular, Eq. (3.2) should generate the change of the field under a hypersurface tilt. From this, we can deduce that the super-Hamiltonian H^Φ must contain a certain part, H^Φ_τ, which is determined solely by the tensor character of the field,

$$H^\Phi=H^\Phi_\dagger+H^\Phi_\tau. \tag{4.4}$$

The remaining part of the super-Hamiltonian, $H^\Phi_\dagger$, generates the evolution of the field under a hypersurface translation, and its detailed structure depends on the specific Lagrangian which governs the field dynamics.

However, there is at least something which we can say at the kinematical level about the translation part $H^\Phi_\dagger$ of the super-Hamiltonian H^Φ: $H^\Phi_\dagger(x)$ must be purely local in the field momenta $\pi^\perp(x)$, $\pi^a(x)$ [i.e., it must be an algebraic function of $\pi^\perp(x)$ and $\pi^a(x)$] in order that the changes $\delta_\dagger\phi_\perp$ and $\delta_\dagger\phi_a$ or the field projections under a hypersurface translation be proportional to the lapse function. Similarly, we can see that the tilt part H^Φ_τ of the super-Hamiltonian must contain the first space derivatives of the field momenta, in order to generate the first derivatives of the lapse functions in the tilt formulas, like (3.1) or (3.16). In fact, we will now show that H^Φ_τ for any tensor field may be chosen as a space divergence of a certain space vector density, which is bilinear in the field variables and the field momenta.

Indeed, comparing the tilt displacement formulas (3.1) for a covector field ϕ_α with the evolution equations (4.2),

$$\begin{aligned}
\delta_\tau\phi_\perp(x)&=-\phi^a(x)N_{,a}(x)=[\phi_\perp(x),H^\Phi_{\tau x'}]N^{x'}\\
&=\frac{\delta H^\Phi_{\tau x'}}{\delta\pi^\perp(x)}N^{x'},\\
\delta_\tau\phi_a(x)&=\phi_\perp(x)N_{,a}(x)=[\phi_a(x),H^\Phi_{\tau x'}]N^{x'}\\
&=\frac{\delta H^\Phi_{\tau x'}}{\delta\pi^a(x)}N^{x'},
\end{aligned}$$

we see that the tilt super-Hamiltonian H^Φ_τ must be equal to

$$H^\Phi_\tau=(\phi^a\pi^\perp-\epsilon\phi_\perp\pi^a)_{,a}. \tag{4.5}$$

Similarly, for a tensor field $\phi_{\alpha\beta}$, we compare Eq. (3.16) with Eq. (4.2), and get

$$\begin{aligned}
H^\Phi_\tau=[&(\phi_\perp{}^a+\phi^a{}_\perp)\pi^{\perp\perp}+\phi_b{}^a\pi^{b\perp}-\epsilon\phi_{\perp\perp}\pi^{a\perp}\\
&+\phi^a{}_b\pi^{\perp b}-\epsilon\phi_{\perp\perp}\pi^{\perp a}-\epsilon\phi_{\perp b}\pi^{ab}-\epsilon\phi_{b\perp}\pi^{ba}]_{,a}.
\end{aligned} \tag{4.6}$$

For higher order tensor fields, the formulas soon get messy, but the general statement that H^Φ_τ is a divergence of a vector density bilinear in the field variables and the field momenta, is preserved.

For two important tensor fields, the tilt super-Hamiltonian vanishes. These are the scalar field ϕ and the metric field $g_{\alpha\beta}$ itself.[6] For the scalar field, the vanishing of H^Φ_τ is obvious, because ϕ is not changed under hypersurface tilts. For the metric field, the vanishing of H^Φ_τ is equally obvious, because $\delta_\tau g_{ab}=0$ and $g_{a\perp}=0$, $g_{\perp\perp}=\epsilon$ are not changed by hypersurface tilts. The super-Hamiltonians of these two field are the only super-Hamiltonians which are purely local in the field momenta. (The special situation which arises for the electromagnetic field due to the additional constraints will be discussed in the following paper).

5. CANONICAL REALIZATION OF TANGENTIAL DEFORMATIONS

The second type of deformations under which the dynamical variables change in a kinematically predictable way are the tangential displacements

$$(\mathbf{N})^{\alpha x}=N^a(x)\,e^\alpha_a(x). \tag{5.1}$$

The field projections $\underset{\sim}{\phi}$ and their conjugate momenta $\underset{\sim}{\pi}$

are hypertensors. Further, because all spacetime indices are projected, $\underset{\sim}{\phi}$ and $\tilde{\pi}$ are spacetime scalars, i.e., hypertensors of the rank $\binom{0}{0;r}$. We have seen that

$$\mathbf{L}_{\tilde{N}} = \delta_{\tilde{N}} - L_{\tilde{N}} \tag{5.2}$$

(see Sec. I.7). The hypertensor condition

$$\mathbf{L}_{\tilde{N}} \underset{\sim}{\phi} = 0 = \mathbf{L}_{\tilde{N}} \tilde{\pi} \tag{5.3}$$

can thus be translated into the equations

$$\delta_{\tilde{N}} \underset{\sim}{\phi} = L_{\tilde{N}} \underset{\sim}{\phi}, \quad \delta_{\tilde{N}} \tilde{\pi} = L_{\tilde{N}} \tilde{\pi}, \tag{5.4}$$

which determine the change of the dynamical variables $\underset{\sim}{\phi}$, $\tilde{\pi}$ under a tangential deformation of the embedding.

In a dynamical theory, the same tangential change is generated by the Poisson bracket with the supermomentum

$$H^{\Phi}{}_{a} = H^{\Phi}{}_{a}(x)[\underset{\sim}{\phi}, \tilde{\pi}]$$

according to the formulas

$$\begin{aligned} \delta_{\tilde{N}} \phi\ldots(x) &= [\phi\ldots(x),\ H^{\Phi}{}_{ax'}] N^{ax'} \\ &= \frac{\delta H^{\Phi}{}_{ax'}}{\delta \pi^{\cdots}(x)} N^{ax'}, \\ \delta_{\tilde{N}} \pi^{\cdots}(x) &= [\pi^{\cdots}(x),\ H^{\Phi}{}_{ax'}] N^{ax'} \\ &= -\frac{\delta H^{\Phi}{}_{ax'}}{\delta \phi\ldots(x)} N^{ax'}. \end{aligned} \tag{5.5}$$

Because the tangential deformations (5.4) and (5.5) must coincide for any $\tilde{N}(x')$, the variational derivatives $\delta H^{\Phi}{}_{a}(x')/\delta\phi\ldots(x)$ and $\delta H^{\Phi}{}_{a}(x')/\delta\pi^{\cdots}(x)$, and with them the supermomentum $H^{\Phi}{}_{a}$ itself, are uniquely determined.

As an example, for the scalar field $\phi(x)$ with its conjugate momentum density $\pi(x)$, we get the equations

$$\begin{aligned} &\frac{\delta H^{\Phi}{}_{ax'}}{\delta\pi(x)} N^{ax'} = L_{\tilde{N}} \phi(x) = \phi_{,a}(x) N^{a}(x), \\ &-\frac{\delta H^{\Phi}{}_{ax'}}{\delta\phi(x)} N^{ax'} = L_{\tilde{N}} \pi(x) = (\pi(x) N^{a}(x))_{,a}. \end{aligned} \tag{5.6}$$

These equations, due to the arbitrariness of $N^{a}(x)$, may be written in the form

$$\begin{aligned} &\frac{\delta H^{\Phi}{}_{a}(x')}{\delta\pi(x)} = \phi_{,a}(x)\, \delta(x, x'), \\ &\frac{\delta H^{\Phi}{}_{a}(x')}{\delta\phi(x)} = -\pi_{,a}(x)\, \delta(x, x') - \pi(x)\, \delta_{,a}(x, x'), \end{aligned} \tag{5.7}$$

and integrated into

$$H^{\Phi}{}_{a}(x) = \pi(x)\, \phi_{,a}(x). \tag{5.8}$$

A similar analysis of the vector case leads to the system of equations

$$\begin{aligned} &\frac{\delta H^{\Phi}{}_{bx'}}{\delta\pi^{\perp}(x)} N^{bx'} = L_{\tilde{N}} \phi_{\perp}(x) = \phi_{\perp,b}(x) N^{b}(x), \\ &-\frac{\delta H^{\Phi}{}_{bx'}}{\delta\phi_{\perp}(x)} N^{bx'} = L_{\tilde{N}} \pi^{\perp}(x) = (\pi^{\perp}(x) N^{b}(x))_{,b}, \\ &\frac{\delta H^{\Phi}{}_{bx'}}{\delta\pi^{a}(x)} N^{bx'} = L_{\tilde{N}} \phi_{a}(x) = \phi_{a,b} N^{b} + \phi_{b} N^{b}{}_{,a}, \\ &-\frac{\delta H^{\Phi}{}_{bx'}}{\delta\phi_{a}(x)} N^{bx'} = L_{\tilde{N}} \pi^{a}(x) = (\pi^{a} N^{b})_{,b} - \pi^{b} N^{a}{}_{,b}, \end{aligned} \tag{5.9}$$

which has the unique solution

$$H^{\Phi}{}_{a} = \pi^{\perp} \phi_{\perp,a} - (\phi_{a,b} - \phi_{b,a})\, \pi^{b} - \phi_{a}\, \pi^{b}{}_{,b}. \tag{5.10}$$

Finally, for a tensor field $\phi_{\alpha\beta}$, which is the last field we want to discuss in detail, we get

$$\begin{aligned} H^{\Phi}{}_{a} = {} & \pi^{\perp\perp} \phi_{\perp\perp,a} \\ & + \pi^{b\perp} \phi_{b\perp,a} - (\pi^{b\perp} \phi_{a\perp})_{,b} \\ & + \pi^{\perp b} \phi_{\perp b,a} - (\pi^{\perp b} \phi_{\perp a})_{,b} \\ & + \pi^{bc} \phi_{bc,a} - (\pi^{bc} \phi_{ac})_{,b} - (\pi^{bc} \phi_{ba})_{,c}. \end{aligned} \tag{5.11}$$

Following the sequence of formulas (5.8), (5.10), (5.11), we can see how to build the supermomentum for the higher-rank tensor fields. The scalar projection $\phi_{\perp}$ of the covector field ϕ_{α} contributes to the covector field supermomentum (5.10), and the scalar projection $\phi_{\perp\perp}$ of $\phi_{\alpha\beta}$ contributes to the tensor field supermomentum (5.11) by the terms which have the form of the scalar field ϕ supermomentum (5.8). Similarly, the covector projections $\phi_{\perp a}$ and $\phi_{a\perp}$ of the tensor field $\phi_{\alpha\beta}$ contribute to the tensor field supermomentum by the terms } which have the form of the covector field supermomentum (5.10). The construction of the successive terms of higher and higher rank also becomes apparent when we follow the consecutive lines in the expression (5.11).

Specializing the supermomentum (5.11) to antisymmetrical tensors (two-forms), we get

$$H^{\Phi}{}_{a} = \pi^{bc} \phi_{bc,a} - 2(\pi^{bc} \phi_{ac})_{,b}. \tag{5.12}$$

Similarly, specializing it to symmetrical tensors, we get

$$\begin{aligned} H^{\Phi}{}_{a} = {} & \pi^{\perp\perp} \phi_{\perp\perp,a} + 2\pi^{b\perp} (\phi_{b\perp,a} - \phi_{a\perp,b}) \\ & - 2\phi_{a\perp} \pi^{b\perp}{}_{,b} + \pi^{bc} \phi_{bc,a} - 2(\pi^{bc} \phi_{ac})_{,b}. \end{aligned} \tag{5.13}$$

In particular, for the metric tensor $g_{\alpha\beta}$ in the role of $\phi_{\alpha\beta}$, the projections $g_{\perp\perp} = \epsilon$ and $g_{a\perp} = 0$ give the vanishing contribution, and the expression (5.13) reduces to

$$H^{g}{}_{a} = -2(\pi^{bc} g_{ac})_{,b} + \pi^{bc} g_{bc,a} = -2\pi^{b}_{a|b}, \tag{5.14}$$

the well-known expression for the gravitational supermomentum.

Summarizing the results of the last two sections, we can say that the tensor field supermomentum and the tilt part of its super-Hamiltonian (the part which is nonlocal in the field momenta) are completely determined from the purely kinematical considerations. It is only the translational part of the super-Hamiltonian (the part which is local in the field momenta), which is truly dynamical and requires the knowledge of the field Lagrangian for its determination.

6. PROJECTIONS OF THE RIEMANN TENSOR

The Riemann curvature tensors $\underset{\sim}{R}$ and $\underline{R}$ in space and in spacetime are defined by the commutation relations of the covariant derivatives | and ; ,

$$\phi_{\alpha;[\beta\gamma]} = \tfrac{1}{2} \phi^{\delta}\, {}^{4}R_{\delta\alpha\beta\gamma}, \tag{6.1}$$

$$\phi_{a|[bc]} = \tfrac{1}{2} \phi^{d} R_{dabc}. \tag{6.2}$$

Projecting Eq. (6.1), we find the three algebraically in-

dependent projections, ${}^4R_{abcd}$, ${}^4R_{\perp abc}$, and ${}^4R_{\perp a\perp b}$, of the spacetime curvature tensor $\underline{R}$.

The abc projection of Eq. (6.1) gives

$$(\phi_{a;[b})_{;c]} = \tfrac{1}{2}\epsilon\phi^{\perp}\,{}^4R_{\perp abc} + \phi^d\,{}^4R_{dabc}. \tag{6.3}$$

Using first Eqs. (2.7) and then Eq. (2.5), we get

$$\begin{aligned}(\phi_{a;[b})_{;c]} &= (\phi_{a;[b})_{|c]} - K_{a[c}\phi_{\perp;b]} - K_{[bc]}\phi_{a;\perp} \\ &= (\phi_{a[b} - \phi_{\perp}K_{a[b})_{|c]} - K_{a[c}(\phi_{\perp|b]} + \epsilon K_{b]d}\,\phi^d) \\ &= \tfrac{1}{2}\phi^d R_{dabc} - \phi^{\perp}K_{a[b|c]} + \epsilon\phi^d K_{a[b}K_{c]d}.\end{aligned} \tag{6.4}$$

Comparing Eqs. (6.3) and (6.4), we get the Gauss–Codazzi equations

$$\begin{aligned}{}^4R_{abcd} &= -2\epsilon K_{a[c}K_{bd]} + R_{abcd}, \\ {}^4R_{\perp abc} &= -2\epsilon K_{a[b|c]}.\end{aligned} \tag{6.5}$$

To get the ${}^4R_{\perp a\perp b}$ projection is slightly more difficult. We start from the $\perp b\perp$ projection of Eq. (6.1),

$$N\phi_{\perp;[b\perp]} = \tfrac{1}{2}N\phi^d\,{}^4R_{d\perp b\perp}. \tag{6.6}$$

Substituting here for $(\phi_{\perp;b})_{;\perp}$ and $(\phi_{\perp;\perp})_{;b}$ from Eqs. (2.7) and (2.8), we get

$$\begin{aligned}N\phi^d\,{}^4R_{d\perp b\perp} &= \epsilon\,\delta_N(\phi_{\perp;b}) \\ &\quad - (N\phi_{\perp;\perp})_{|b} - \epsilon K_b^d(N\phi_{d;\perp}) + \epsilon\phi^d{}_{;b}\,N_{,d}.\end{aligned} \tag{6.7}$$

Again, we substitute here for $\phi_{\perp;b}$, $N\phi_{\perp;\perp}$, $N\phi_{d;\perp}$, and $\phi^d{}_{;b}$ from Eqs. (2.5), (2.6). Using Eq. (I.9.8),

$$\delta_N g^{ab} = 2NK^{ab}, \tag{6.8}$$

and realizing that $\delta_N(\phi_{\perp,b}) = (\delta_N\phi_{\perp})_{,b}$, most of the terms cancel and we get

$$\delta_N K_{ab} = ({}^4R_{\perp a\perp b} - K_a^c K_{cb})N + \epsilon N_{|ab}. \tag{6.9}$$

This is a counterpart of Eq. (I.9.7),

$$\delta_N g_{ab} = -2K_{ab}N, \tag{6.10}$$

for the normal change of the intrinsic geometry. Equations (6.9) and (6.10) for g_{ab} and K_{ab}, however, do not close, because we still ought to know ${}^4R_{\perp a\perp b}$ before predicting what extrinsic curvature K_{ab} we shall find on the deformed hypersurface. In Sec. 8, we shall rewrite Eq. (6.9) into a new form, connecting it with the G_{ab} projection of the Einstein tensor $G_{\alpha\beta}$.

7. ON CANONICAL REALIZATION OF GEOMETROKINEMATICS

The normal changes of g_{ab} and K_{ab} in an arbitrary spacetime are governed by Eqs. (6.10) and (6.9). Any dynamical theory which aims at reconstructing the spacetime by a canonical evolution of the space geometry g must respect the kinematical relations (6.9) and (6.10). In such a theory, the normal change of an arbitrary dynamical variable $F[g_{ab}, \pi^{ab}]$ is determined by its Poisson bracket with the geometrodynamical super-Hamiltonian

$$H^{\mathcal{E}} = H^{\mathcal{E}}(x)[g_{ab}, \pi^{ab}], \tag{7.1}$$

which is a functional of the metric g_{ab} and its conjugate momentum π^{ab}, according to the formula

$$\delta_N F = [F, H^{\mathcal{E}}{}_{x'}]\,N^{x'}. \tag{7.2}$$

As in the canonical realization of hypersurface tilts, nothing further is known about the nature of the gravitational momentum $\pi^{ab}(x)$, except that it is canonically conjugate to the metric $g_{ab}(x)$, satisfying the Poisson bracket relation

$$[g_{ab}(x), \pi^{cd}(x')] = \delta_{ab}^{cd}\,\delta(x, x'). \tag{7.3}$$

In particular, the relation between π^{ab} and K_{ab} remains unknown; we assume only that K_{ab} is a dynamical variable, expressible as a functional of the conjugate canonical variables g_{ab} and π^{ab},

$$K_{ab} = K_{ab}(x)[g_{cd}, \pi^{cd}]. \tag{7.4}$$

The normal change of K_{ab} is thus given by Eq. (7.2),

$$\delta_N K_{ab}(x) = [K_{ab}(x), H^{\mathcal{E}}{}_{x'}]N^{x'}. \tag{7.5}$$

Comparing this equation with the kinematical relation (6.9), we see that

$$[K_{ab}(x), H^{\mathcal{E}}(x')] = F_{ab}(x)\,\delta(x, x') + \epsilon\delta_{|ab}(x, x'), \tag{7.6}$$

where we have written F_{ab} in place of the coefficient of the lapse function N in Eq. (6.9).

However, according to the kinematical relation (6.10) and the evolution equation (7.2) for the metric $g_{ab}, K_{ab}(x)$ itself is expressible in the form

$$\begin{aligned}-2K_{ab}(x)\,\delta(x, x'') &= \delta_{\perp x''}\,g_{ab}(x) \\ &= [g_{ab}(x), H^{\mathcal{E}}(x'')] = \frac{\delta H^{\mathcal{E}}(x'')}{\delta\pi^{ab}(x)}.\end{aligned} \tag{7.7}$$

Bringing Eqs. (7.6) and (7.7) together, we get

$$\begin{aligned}\left[\frac{\delta H^{\mathcal{E}}(x)}{\delta\pi^{ab}(x'')}, H^{\mathcal{E}}(x')\right] &= -2F_{ab}(x)\,\delta(x, x')\,\delta(x, x'') \\ &\quad - 2\epsilon\delta(x, x'')\,\delta_{|ab}(x, x').\end{aligned} \tag{7.8}$$

A similar equation can be written down, with the points x and x' interchanged. If we subtract it from Eq. (7.8), the unknown term F_{ab} drops out, due to the symmetry of the $\delta(x, x')$ function, and the left-hand side of the resulting equation becomes the variational derivative of $[H(x), H(x')]$, due to the antisymmetry of this Poisson bracket. We thus get

$$\begin{aligned}&\frac{\delta}{\delta\pi^{ab}(x'')}[H^{\mathcal{E}}(x), H^{\mathcal{E}}(x')] \\ &\quad = -2\epsilon\delta(x, x'')\,\delta_{|ab}(x, x') - (x \longleftrightarrow x').\end{aligned} \tag{7.9}$$

The last equation can be functionally integrated with respect to π^{ab}, giving

$$\begin{aligned}[H^{\mathcal{E}}(x), H^{\mathcal{E}}(x')] &= -2\epsilon\pi^{ab}(x)\,\delta_{|ab}(x, x') - (x \longleftrightarrow x') \\ &\quad + f(x, x')[g_{ab}].\end{aligned} \tag{7.10}$$

Here, $f(x, x')[g_{ab}]$ is an arbitrary functional of the metric, antisymmetrical in the points x and x', which plays the role of the constant of integration.

We can introduce the gravitational supermomentum (5.13) into Eq. (7.10), noting the identity

$$H^{\mathcal{E}\,a}(x)\,\delta_{,a}(x, x') - (x \longleftrightarrow x') = 2\pi^{ab}(x)\,\delta_{|ab}(x, x') - (x \longleftrightarrow x'). \tag{7.11}$$

Equation (7.10) then takes the final form

$$[H^g(x), H^g(x')] = -\epsilon H^{g\,a}(x)\,\delta_{,a}(x,x') - (x \leftrightarrow x') + f(x,x')[g_{ab}]. \tag{7.12}$$

Studying the evolution of g and $\underset{\sim}{K}$ up to the second order in the lapse function $\tilde{N}$ (which we shall not do here), we can actually prove that $f(x,x')[g_{ab}]$ must be put equal to zero. The Poisson brackets between $H^g(x)$ and $H^g(x')$ then close exactly in the same way as the Lie brackets (I. 6.17) between the $\mathcal{E}$-vectors $\delta_{\perp x}$ and $\delta_{\perp x'}$ of the normal $\mathcal{E}$-basis, $H^g{}_a(x)$ playing the role of δ_{ax}. We can prove then that the only super-Hamiltonian $H^g(x)[g,\tilde{\pi}]$ which solves Eq. (7.12) (with $f=0$) is the standard ADM super-Hamiltonian (8.12) with an extra cosmological term $2\lambda g^{1/2}$. In this way, the Einstein's theory can be regarded as the unique canonical realization of geometrokinematics without sources.[2]

8. GEOMETROKINEMATICS IN FULL

Working out the different projections of the spacetime Ricci tensor, we get

$$\begin{aligned} {}^4R_{\perp\perp} &= {}^4R^c{}_{\perp c\perp}, \\ {}^4R_{a\perp} &= {}^4R^c{}_{ac\perp}, \\ {}^4R_{ab} &= {}^4R^c{}_{acb} + \epsilon\,{}^4R_{\perp a\perp b}. \end{aligned} \tag{8.1}$$

Similarly, we get the spacetime scalar curvature

$${}^4R = {}^4R^{cd}{}_{cd} + 2\epsilon\,{}^4R^c{}_{\perp c\perp}. \tag{8.2}$$

From here and the projection formulas of Sec. 6, we can evaluate the projections of the spacetime Einstein's tensor $G_{\alpha\beta}$;

$$N\,{}^4R = 2\epsilon\delta_N K - \epsilon(K_{ab}K^{ab} + K^2)N + RN - 2g^{ab}N_{|ab}, \tag{8.3}$$

$$G_{\perp\perp} = -\tfrac{1}{2}\epsilon\,{}^4R^{cd}{}_{cd} = -\tfrac{1}{2}(K_{ab}K^{ab} - K^2 + \epsilon R), \tag{8.4}$$

$$G_{\perp a} = {}^4R^c{}_{ac\perp} = -\epsilon(K^b_a - K\delta^b_a)_{|b}, \tag{8.5}$$

$$\begin{aligned} NG_{ab} &= N({}^4R^c{}_{acb} + \epsilon\,{}^4R_{\perp a\perp b} - \tfrac{1}{2}\,{}^4R\,g_{ab}) = \epsilon\,\delta_N(K_{ab} - Kg_{ab}) \\ &\quad + \epsilon(2K^c_a K_{cb} - 3KK_{ab} + \tfrac{1}{2}(K_{cd}K^{cd} + K^2)g_{ab})N \\ &\quad - (N_{|ab} - g^{cd}N_{|cd}g_{ab}) + {}^3G_{ab}N. \end{aligned} \tag{8.6}$$

In Eq. (8.6), ${}^3G_{ab}$ is the Einstein's tensor of $(m,\underset{\sim}{g})$.

At this stage, it is useful to introduce the gravitational momentum

$$\pi^{ab} \equiv \epsilon g^{1/2}(K^{ab} - Kg^{ab}). \tag{8.7}$$

The relation between the extrinsic curvature K_{ab} and the gravitational momentum π^{ab} can be written by means of the DeWitt's "supermetric" $G^{ab\,cd}$,

$$\pi^{ab} = \epsilon\,G^{ab\,cd}K_{cd}, \qquad K_{ab} = \epsilon\,G_{ab\,cd}\pi^{cd}, \tag{8.8}$$

with

$$\begin{aligned} G^{ab\,cd} &\equiv \tfrac{1}{2}g^{1/2}(2g^{ab}g^{cd} - g^{ac}g^{bd} - g^{bc}g^{ad}), \\ G_{ab\,cd} &= \tfrac{1}{2}g^{-1/2}(g_{ac}g_{bd} + g_{ad}g_{bc} - g_{ab}g_{cd}), \end{aligned} \tag{8.9}$$

$$\begin{aligned} &G_{ab\,cd} = G_{ba\,cd} = G_{ab\,dc} = G_{cd\,ab}, \\ &G_{ab\,cd}G^{cd\,ef} = \delta^{ef}_{ab} \equiv \tfrac{1}{2}(\delta^e_a\delta^f_b + \delta^f_a\delta^e_b). \end{aligned} \tag{8.10}$$

Equations (8.3) and (8.4) are then easily expressed in terms of π^{ab},

$$\begin{aligned} \underset{\sim}{G}_{\perp\perp} &\equiv g^{1/2}G_{\perp\perp} = \tfrac{1}{2}\epsilon H^g, \\ \underset{\sim}{G}_{\perp a} &\equiv g^{1/2}G_{\perp a} = \tfrac{1}{2}H^g{}_a; \end{aligned} \tag{8.11}$$

H^g is the well-known gravitational super-Hamiltonian (with the cosmological term put equal to zero),

$$H^g \equiv -\epsilon\,G_{ab\,cd}\pi^{ab}\pi^{cd} - g^{1/2}R, \tag{8.12}$$

and $H^g{}_a$ is the gravitational supermomentum

$$H^g{}_a \equiv -2\pi^b_{a|b}, \tag{8.13}$$

which we have already encountered in Eq. (5.14).

Equation (8.6) can also be written in terms of the gravitational momentum, after we raise the indices a and b. Because

$$\delta_N(g^{1/2}g^{ac}g^{bd}) = g^{1/2}(-Kg^{ac}g^{bd} + 2K^{ac}g^{bd} + 2K^{bd}g^{ac}), \tag{8.14}$$

we get

$$\begin{aligned} \delta_N\pi^{ab} &= -\epsilon g^{1/2}(\pi\pi^{ab} - 2\pi^{ac}\pi^b_c + \tfrac{1}{2}(\pi^{cd}\pi_{cd} - \tfrac{1}{2}\pi^2)g^{ab})N \\ &\quad + (N^{|ab} - g^{cd}N_{|cd}g^{ab}) + (G^{ab} - {}^3G^{ab})N. \end{aligned} \tag{8.15}$$

The quadratic combination of π^{ab} on the right-hand side of Eq. (8.15) is equal to

$$\epsilon\frac{\partial}{\partial g_{ab}}(G_{cd\,ef}\pi^{cd}\pi^{ef}). \tag{8.16}$$

Further,

$$\frac{\delta}{\delta g_{ab}(x)}(g^{1/2}R)_{x'}N^{x'} = (N^{|ab} - g^{cd}N_{|cd}g^{ab}) - {}^3G^{ab}N. \tag{8.17}$$

If we consider π^{ab} as the canonical momentum conjugate to g_{ab},

$$[g_{ab}(x), \pi^{cd}(x')] = \delta^{cd}_{ab}\,\delta(x,x'), \tag{8.18}$$

we can write Eq. (8.15) in a very compact form

$$\delta_N\pi^{ab}(x) = [\pi^{ab}(x), H^g{}_{x'}]N^{x'} + G^{ab}N, \tag{8.19}$$

where H^g is the super-Hamiltonian introduced by Eq. (8.12). Similarly,

$$\delta_N g_{ab}(x) = [g_{ab}(x), H^g{}_{x'}]N^{x'}. \tag{8.20}$$

Equations (8.19) and (8.20) hold in an arbitrary Riemannian spacetime and are thus kinematical equations. Note that Einstein's law was never used in their derivation. In effect, Eqs. (8.19) and (8.20) are just a fancy way of writing down the kinematical equations (6.9) and (6.10) Again, Eqs. (8.19) and (8.20) are not closed in the variables g_{ab} and π^{ab}, because Eq. (8.19) contains a source term $G^{ab}N$, which is proportional to the tangential projection of the spacetime Einstein's tensor. While the $G_{\perp\perp}$ and $G_{\perp a}$ projections of the Einstein's tensor are determined by the initial data—the intrinsic geometry g_{ab} and the extrinsic curvature K_{ab} (or the gravitational momentum π^{ab})—by Eqs. (8.11)–(8.13), the tangential projection G_{ab} is not determined by the initial data. In addition to g_{ab} and π^{ab}, G^{ab} must be specified on an initial hypersurface before

the gravitational momentum can be calculated on a deformed hypersurface from Eq. (8.19). If we want to proceed yet to another hypersurface, we must again specify the projection G^{ab} on the deformed hypersurface, and so on at each successive step. It must be so, because otherwise the initial data on an initial hypersurface would determine the whole spacetime geometry by themselves. This is clearly impossible, because spacetime geometry is arbitrary at this stage and may thus be freely readjusted in the regions away from the initial hypersurface.

If we subject the spacetime to Einstein's law of gravitation, G^{ab} becomes proportional to the stress tensor as measured by the family of observers who move in the normal direction to the hypersurface. In a vacuum Einstein's spacetime, $G^{ab}=0$ and the initial data g_{ab}, π^{ab} must satisfy the constraints $H^g=0=H^g{}_a$. Equations (8.19) and (8.20) close in the variables g_{ab}, π^{ab}, becoming canonical equations generated by the Hamiltonian $H^g{}_{x'}N^{x'}$. In this section, we have derived Eqs. (8.19) and (8.20) by the direct projection of the Einstein's law. In the final paper of this series, we will discuss how to get these dynamical equations from an action principle.

9. CONSERVATION LAWS

External sources, like the four-current J^α in electrodynamics or the energy–momentum tensor $T^{\alpha\beta}$ in the Einstein's theory of gravitation, often obey the conservation laws in the form of spacetime divergence equations,

$$J^\alpha{}_{;\alpha}=0, \tag{9.1}$$

$$T_\alpha{}^\beta{}_{;\beta}=0. \tag{9.2}$$

In hypersurface dynamics, the conservation laws tell us how the charge density $\underset{\sim}{J}{}^\perp\equiv g^{1/2}J^\perp$, the energy density $\underset{\sim}{T}{}^{\perp\perp}\equiv g^{1/2}T^{\perp\perp}$, and the momentum density $\underset{\sim}{T}{}^{\perp a}\equiv g^{1/2}T^{\perp a}$, measured by the family of observers moving perpendicular to a hypersurface, change when we pass to another hypersurface. In other words, the conservation laws are to be written as restrictions on the normal changes δ_N of the projections $\underset{\sim}{J}{}^\perp$, $\underset{\sim}{T}{}^{\perp\perp}$, and $\underset{\sim}{T}{}^{\perp a}$. To do that, we project Eqs. (9.1) and (9.2), writing them in the form

$$Ng^{1/2}J^\alpha{}_{;\alpha}=Ng^{1/2}g^{ab}J_{a;b}+\epsilon Ng^{1/2}J_{\perp;\perp}=0, \tag{9.3}$$

and

$$\begin{aligned} Ng^{1/2}T_\perp{}^\beta{}_{;\beta}&=Ng^{1/2}g^{ab}T_{\perp a;b}+\epsilon Ng^{1/2}T_{\perp\perp;\perp}=0,\\ Ng^{1/2}T_a{}^\beta{}_{;\beta}&=Ng^{1/2}g^{bc}T_{ab;c}+\epsilon Ng^{1/2}T_{a\perp;\perp}=0. \end{aligned} \tag{9.4}$$

We then use the formulas (2.5)–(2.8) for the projections of the spacetime covariant derivatives, and Eq. (I.9.9),

$$\delta_N g^{1/2}=-g^{1/2}KN, \tag{9.5}$$

getting

$$Ng^{1/2}J^\alpha{}_{;\alpha}=\delta_N\underset{\sim}{J}{}_\perp+(N\underset{\sim}{J}{}^a)_{,a}=0, \tag{9.6}$$

and

$$Ng^{1/2}T_\perp{}^\beta{}_{;\beta}=\delta_N\underset{\sim}{T}{}_{\perp\perp}+(N\underset{\sim}{T}{}_\perp{}^b)_{,b}+\underset{\sim}{T}{}^b{}_\perp N_{|b}+\epsilon K^{ab}\underset{\sim}{T}{}_{ab}N=0, \tag{9.7}$$

$$Ng^{1/2}T_a{}^\beta{}_{;\beta}=\delta_N\underset{\sim}{T}{}_{a\perp}+(N\underset{\sim}{T}{}_a{}^b)_{|b}-\epsilon\underset{\sim}{T}{}_{\perp\perp}N_{|a}+NK^b_a(\underset{\sim}{T}{}_{b\perp}-\underset{\sim}{T}{}_{\perp b})=0. \tag{9.8}$$

Note that the extrinsic curvature does not enter into Eq. (9.6), corresponding to the fact that the conservation law $J^\alpha{}_{;\alpha}=0$ can be written in the form

$$(N(-{}^4g)^{1/2}J^\alpha)_{,\alpha}=0,$$

using only the partial derivatives of the spacetime vector density $(-{}^4g)^{1/2}J^\alpha$. For a symmetrical tensor, the last term in Eq. (9.8) vanishes and the extrinsic curvature thus disappears from the momentum conservation law as well. However, the extrinsic curvature remains in the conservation law (9.7) for the energy.

A number of terms in Eqs. (9.6)–(9.8) can be easily understood if we realize that $\{J_\perp, J_a\}$ and $\{T_{\perp\perp}, T_{\perp a}, T_{a\perp}, T_{ab}\}$, like the projections of any spacetime vector J^α and any spacetime tensor $T^{\alpha\beta}$, must behave in a definite way under hypersurface tilts. Indeed, for hypersurface tilts, Eqs. (9.6)–(9.8) pass into the already known equations (3.1) and (3.16). The real information about the charge conservation is carried only in the translation parts of Eqs. (9.6)–(9.8), namely,

$$\delta_\tau\underset{\sim}{J}{}_\perp=-\underset{\sim}{J}{}^a{}_{,a}N, \tag{9.9}$$

$$\delta_\tau\underset{\sim}{T}{}_{\perp\perp}=-(\underset{\sim}{T}{}_\perp{}^b{}_{,b}+\epsilon K^{ab}\underset{\sim}{T}{}_{ab})N, \tag{9.10}$$

$$\delta_\tau\underset{\sim}{T}{}_{a\perp}=-\underset{\sim}{T}{}_a{}^b{}_{|b}N. \tag{9.11}$$

Using the arbitrariness of the lapse function, Eqs. (9.6)–(9.8) can also be written in the form ($T^{\alpha\beta}$ being taken symmetrical),

$$\begin{aligned} g^{1/2}J^\alpha{}_{;\alpha}\,\delta(x,x')&=\delta_{\perp x'}\underset{\sim}{J}{}_\perp(x)+\underset{\sim}{J}{}^a\delta_{,a}(x,x')\\ &\quad+\underset{\sim}{J}{}^a{}_{,a}\,\delta(x,x')=0, \end{aligned} \tag{9.12}$$

$$\begin{aligned} g^{1/2}T_\perp{}^\beta{}_{;\beta}\,\delta(x,x')&=\delta_{\perp x'}\underset{\sim}{T}{}_{\perp\perp}(x)+\underset{\sim}{T}{}_\perp{}^b{}_{,b}\,\delta(x,x')\\ &\quad+2\underset{\sim}{T}{}_\perp{}^b(x)\,\delta_{,b}(x,x')\\ &\quad+\epsilon K^{ab}\underset{\sim}{T}{}_{ab}\,\delta(x,x')=0, \end{aligned} \tag{9.13}$$

$$\begin{aligned} g^{1/2}T_a{}^\beta{}_{;\beta}\,\delta(x,x')&=\delta_{\perp x'}\underset{\sim}{T}{}_{a\perp}(x)+\underset{\sim}{T}{}_a^b{}_{|b}\,\delta(x,x')\\ &\quad+\underset{\sim}{T}{}_a^b(x)\,\delta_{,b}(x,x')\\ &\quad-\epsilon\underset{\sim}{T}{}_{\perp\perp}(x)\,\delta_{,a}(x,x')=0. \end{aligned} \tag{9.14}$$

10. BIANCHI IDENTITIES

The Einstein's tensor $G^{\alpha\beta}$ satisfies the divergence equation (9.2) identically, by virtue of the contracted Bianchi identities. To express the contracted Bianchi identities in hypersurface language, apply Eq. (9.13) to the Einstein's tensor, using the definitions (8.11) of the super-Hamiltonian and supermomentum,

$$\begin{aligned} 0\equiv 2\epsilon g^{1/2}G_\perp{}^\beta{}_{;\beta}\,\delta(x,x')&=\delta_{\perp x'}H^g(x)+\epsilon H^g{}^a{}_{|a}(x)\,\delta(x,x')\\ &\quad+2\epsilon H^g{}^a(x)\,\delta_{,a}(x,x')+2G^{ab}(x)\,K_{ab}(x)\,\delta(x,x'). \end{aligned} \tag{10.1}$$

On the other hand, $H^g(x)$ is a functional of g_{ab} and π^{ab}, and its normal change can thus be evaluated from Eqs. (8.19) and (8.20),

$$\delta_{\perp x'}H^g(x)=[H^g(x),H^g(x')]+[g_{abx'},H^g(x)]\,G^{abx'}. \tag{10.2}$$

Further,

$$[g_{ab}(x'), H^{\varepsilon}(x)] = \delta_{\perp x} g_{ab}(x') = -2K_{ab}(x')\,\delta(x', x). \quad (10.3)$$

Putting Eqs. (10.1)–(10.3) together, we get

$$0 \equiv -2g^{1/2} G^{\perp\beta}{}_{;\beta}\,\delta(x, x') = [H^{\varepsilon}(x), H^{\varepsilon}(x')] + \epsilon H^{\varepsilon\,a}{}_{|a}(x)\,\delta(x, x') + 2\epsilon H^{\varepsilon\,a}(x)\,\delta_{,a}(x, x'). \quad (10.4)$$

This is the well-known[7] closing relation for the gravitational super-Hamiltonians. It is equivalent to Eq. (7.12), with $f = 0$.

Similarly, start from Eq. (9.8), substituting into it the projections (8.11) of the Einstein's tensor,

$$0 \equiv 2g^{1/2} G_a{}^{\beta}{}_{;\beta} N = \delta_N H^{\varepsilon}{}_a + 2(NG_a{}^b)_{|b} - H^{\varepsilon} N_{|a}. \quad (10.5)$$

Evaluate again the normal change of $H^{\varepsilon}{}_a$ from Eqs. (8.19) and (8.20), getting

$$\delta_N H^{\varepsilon}{}_a(x) = [H^{\varepsilon}{}_a(x), H^{\varepsilon}{}_{x'}] N^{x'} + (G^{bc}N)^{x'} [g_{bcx'}, H^{\varepsilon}{}_a(x)]. \quad (10.6)$$

Because

$$[g_{bc}(x'), H^{\varepsilon}{}_{ax}] N^{ax} = \delta_{\tilde{N}} g_{bc}(x') = 2N_{(b|c)}, \quad (10.7)$$

we can integrate by parts the expression

$$(G^{bc}N)^{x'} [g_{bcx'}, H^{\varepsilon}{}_{ax}] N^{ax} = 2(G^{bc}N)^{x'} N_{b|cx'} = -2(G_b{}^c N)_{|cx'} N^{bx'},$$

and deduce from there the equation

$$(G^{bc}N)^{x'} [g_{bcx'}, H^{\varepsilon}{}_{ax}] = -2(G_a^c N)_{|c}. \quad (10.8)$$

Putting Eqs. (10.5), (10.6), and (10.8) together, we see that

$$0 \equiv -2\epsilon g^{1/2} G_a{}^{\beta}{}_{;\beta} N = [H^{\varepsilon}{}_a(x), H^{\varepsilon}{}_{x'}] N^{x'} - H^{\varepsilon} N_{|a}, \quad (10.9)$$

or

$$0 \equiv 2g^{1/2} G_a{}^{\beta}{}_{;\beta}\,\delta(x, x') = [H^{\varepsilon}{}_a(x), H^{\varepsilon}(x')] - H^{\varepsilon}(x)\,\delta_{,a}(x, x'). \quad (10.10)$$

This is an equally well-known closing relation between the supermomentum $H^{\varepsilon}{}_a(x)$ and the super-Hamiltonian $H^{\varepsilon}(x')$.

*Work supported in part by the National Science Foundation under Grant No. GP-43718X to the University of Utah.

[1]K. Kuchař, J. Math. Phys. **17**, 777 (1976).
[2]S. A. Hojman, K. Kuchař, and C. Teitelboim, Nature 245, 97 (1973); K. Kuchař, J. Math. Phys. 15, 708 (1974); S. A. Hojman, K. Kuchař, and C. Teitelboim, Ann. Phys. (N.Y.), to be published.
[3]C. Teitelboim, 1973 private communication.
[4]P. A. M. Dirac, *Lectures on Quantum Mechanics* (Academic, New York, 1965).
[5]R. Arnowitt, S. Deser, and C. W. Misner, J. Math. Phys. **1**, 434 (1960).
[6]See K. Kuchař, Ref. 2.
[7]See, e.g., B. S. DeWitt, Phys. Rev. **160**, 113 (1967).

Dynamics of tensor fields in hyperspace. III*

Karel Kuchař

Department of Physics, University of Utah, Salt Lake City, Utah 84112
(Received 7 November 1975)

The dynamics of tensor fields with derivative gravitational coupling on a given Riemannian background is formulated as Hamiltonian dynamics of hypersurface projections of these fields propagating in hyperspace. The first-order spacetime action is transformed into an equivalent hypersurface form. The supermomentum and different parts of the super-Hamiltonian are identified with projected pieces of the (symmetrical, canonical, and spin) energy–momentum tensors, and their kinematical and dynamical roles are analyzed. Hypersurface variables are included among the canonical variables, and the resulting first-order generalized Hamiltonian dynamics of hypertensor fields is discussed. The closing relations for the constraint functions in the generalized Hamiltonian dynamics are derived from the foliation independence of the hypersurface action. The elimination of the λ-multipliers, which are characteristic to the first-order theory, is accomplished. The general formalism is specialized to the n-form fields with nonderivative gravitational coupling.

1. INTRODUCTION

Hypersurface dynamics gives the rules according to which the perpendicular and parallel projections of spacetime tensor fields to a spacelike hypersurface change when the hypersurface is deformed through a Riemannian spacetime. Following the program set in our previous papers,[1,2] we visualize the hypersurface dynamics as dynamics of hypertensor fields in hyperspace. Hyperspace is an infinitely dimensional manifold of all spacelike hypersurfaces drawn in a Riemannian spacetime (M, g). It has a rich geometrical structure which we have studied in Ref. 1. The projections of spacetime tensor fields form a fiber of hypertensors over the hypersurface. Hypersurface dynamics tells us how the field point in this fiber moves when the base point follows a curve in hyperspace. Starting from the spacetime action functional, we endow the fiber of hypertensors with the Hamiltonian structure and cast the hypersurface dynamics into a Hamiltonian form.

The idea of hypersurface dynamics of tensor fields on a flat Minkowskian background goes back to Dirac.[3] The hypersurface dynamics of simple tensor fields (scalar and electromagnetic) with nonderivative gravitational coupling was studied in detail.[4] The general features of hypersurface dynamics of tensor fields with derivative gravitational coupling, on the other hand, were never elaborated into a complete scheme. In this paper, we build such a scheme, paying special attention to the correct interpretation of various pieces of the hypersurface Hamiltonian, and to the matching of the kinematical[2] and dynamical aspects of the theory.

We organize the material into twelve sections. First, we remind the reader what is the first-order form of the spacetime field action (Sec. 2) and how it generates the symmetrical, canonical, and spin energy-momentum tensors (Sec. 3). The first-order form of the action is better suited to hypersurface dynamics than the second-order form, because it is directly related to the Hamiltonian formalism, its projections are simpler, and the projected terms are easier to interpret. The detailed knowledge of the construction of the energy-momentum tensors is needed, because the projected parts of these tensors play important and conceptually different roles in hypersurface dynamics. In Sec. 4, we transform the spacetime action S^Φ into an equivalent hypersurface action S^Φ, which is expressed as a functional of the field projections $\perp$ and $\parallel$ to a foliation $e(t)$ of embeddings, and of the foliation itself. The hypersurface action S^Φ of a covector field is worked out in Sec. 5, and cast into the Hamiltonian form. The constituent parts of the field super-Hamiltonian and supermomentum are identified. The algorithm developed in this section is easily generalized to higher-rank tensor fields. (It is applied to the second-rank tensor fields in Sec. 9.) The Hamiltonian dynamics of spacetime hypertensors generated by the hypersurface action is discussed in Sec. 6. The first-order form of the action leads to the appearance of the Lagrange multipliers λ in the hypersurface Lagrangian and to the separation of the Euler equations into Hamilton's equations and λ equations. The mixing of the field momenta and the λ multipliers under hypersurface tilts leads to the discovery of the kinematical role of the tilt super-Hamiltonian $H_\perp^{(\tau)}$. The symplectic structure of the hypertensor phase space P and the many-fingered-time nature of field curves in P are discussed in geometrical terms. In Sec. 7, the embedding e with conjugate energy-momentum densities $\mathfrak{p}$ is included among the canonical variables, and the generalized Hamiltonian dynamics of hypertensor fields on the constraint hypersurface in the generalized phase space ${}^eP$ is developed. In Sec. 8, we show what roles are played by various projections of the energy-momentum tensors in hypersurface dynamics. In particular, we generate all projections, $T^{\perp\perp}$, $T^{\perp b}$, and T^{ab}, of the symmetrical energy-momentum tensor $T^{\alpha\beta}$ directly from the hypersurface Lagrangian. As we have already mentioned, Sec. 9 constructs the hypersurface Lagrangian for second-rank tensor fields. In Sec. 10, we explain the process by which the λ multipliers are eliminated from the formalism on the level of the hypersurface action. We illustrate the process on a covector field satisfying the wave equation. The hypersurface dynamics notably simplifies for the fields with nonderivative gravitational coupling (no derivatives of

the metric tensor $\underline{g}$ appear in the spacetime action). In Sec. 11, we give a systematic treatment of such fields (n-form fields $\underline{\phi} \in \mathcal{T}_{\underline{g}}(\mathcal{M})$, $n = 0, 1, 2, 3$), and study the Proca's and Maxwell's fields as illustrations of the general theory. In Sec. 12, we derive the closing relations between the constraint functions in generalized Hamiltonian dynamics of hypertensor fields from the invariance of the hypersurface action under the change of foliation. These closing relations are complicated by terms involving the λ equations. After the λ multipliers are eliminated, the closing relations assume the standard universal form.

The dynamical interaction between geometry and the tensor fields with derivative gravitational coupling is the subject of the last paper of this series, "Tensor Sources in Geometrodynamics."

2. FIELD ACTION

In this paper, we study the tensor fields $\underline{\phi}$ propagating on a given Riemannian background $(\mathcal{M}, \underline{g})$. For uniformity, we always take the fields $\underline{\phi}$ in a completely covariant form. We assume that the field equations follow from the field action

$$S^{\Phi} = S^{\Phi}[\underline{\phi}] \equiv \int \underline{\eta} L(\underline{\phi}, \nabla\underline{\phi}, \underline{g}), \tag{2.1}$$

which is an integral of the Lagrangian L with respect to the Levi-Civita form $\underline{\eta}$. The Lagrangian is a scalar invariant constructed from the field $\underline{\phi}$, its first covariant derivatives $\nabla\underline{\phi}$, and the metric $\underline{g}$ of the background. In a coordinate basis $\underline{\partial}_\alpha$,

$$L = L(\phi_{\{\alpha\}}, \phi_{\{\alpha\};\beta}, g_{\alpha\beta}), \tag{2.2}$$

where $\{\alpha\}$ stands for the collection $\alpha_1 \cdots \alpha_n$ of indices, and

$$\eta_{\alpha\beta\gamma\delta} = |{}^4g|^{1/2} \delta_{\alpha\beta\gamma\delta}. \tag{2.3}$$

The variation of the field action (2.1) with respect to $\underline{\phi}$ yields the field equations

$$\frac{\delta S^{\Phi}}{\delta \underline{\phi}(X)} = |{}^4g|^{1/2} \left(\frac{\partial L}{\partial \underline{\phi}} - \nabla_\beta \frac{\partial L}{\partial (\nabla_\beta \underline{\phi})} \right) = 0. \tag{2.4}$$

The dependence of the second-order Lagrangian L on the field derivatives $\nabla\underline{\phi}$ may be fairly complicated. In hypersurface dynamics, we are required to project everything into the normal and tangential directions to spacelike hypersurfaces, and cast the action into a Hamiltonian form. The projections are much easier to handle if the covariant derivatives enter the action in a standard linear way which is directly connected with the Hamiltonian formalism. It is thus useful to bring the action into such a "first-order form," at the price of introducing supplementary variables $\overline{\lambda}$ which form a contravariant tensor of rank $n+1$. Certain projections of this contravariant tensor are then identified with the field momenta, while other projections stay in the formalism as Lagrange multipliers. The interpretation of the hypersurface dynamics is most easily carried through in this form, the Lagrange multipliers being eliminated only in the last step.

We pass from the "second-order form" (2.1) of the action to the "first-order form" by putting

$$\overline{\lambda} = \frac{\partial L}{\partial (\nabla \underline{\phi})}, \qquad \lambda^{\{\alpha\}\beta} = \frac{\partial L}{\partial \phi_{\{\alpha\};\beta}} \tag{2.5}$$

and performing the Legendre dual transformation

$$\underline{\phi}, \quad \nabla\underline{\phi}, \quad L \to \underline{\phi}, \quad \overline{\lambda}, \quad \Lambda. \tag{2.6}$$

Assuming that Eq. (2.5) is invertible with respect to $\nabla\underline{\phi}$,

$$\nabla\underline{\phi} = \nabla\underline{\phi}(\underline{\phi}, \overline{\lambda}, \underline{g}), \tag{2.7}$$

we introduce the function Λ by the equation

$$\begin{aligned} \Lambda &= [\overline{\lambda} \lrcorner \nabla\underline{\phi} - L]_{\nabla\underline{\phi} = \nabla\underline{\phi}(\underline{\phi}, \overline{\lambda}, \underline{g})} \\ &= [\lambda^{\{\alpha\}\beta} \phi_{\{\alpha\};\beta} - L]_{\phi_{\{\alpha\};\beta} = \phi_{\{\alpha\};\beta}(\phi_{\{\gamma\}}, \lambda^{\{\gamma\}\delta}, g_{\gamma\delta})}, \end{aligned} \tag{2.8}$$

and express the action (2.1) in terms of the new variables $\underline{\phi}, \overline{\lambda}$,

$$S^{\Phi} = S^{\Phi}[\underline{\phi}, \overline{\lambda}] = \int_{\mathcal{M}} \underline{\eta} (\overline{\lambda} \lrcorner \nabla\underline{\phi} - \Lambda). \tag{2.9}$$

This is the first-order form of the action. The scalar invariant $\Lambda(\underline{\phi}, \overline{\lambda}, \underline{g})$ completely characterizes the dynamical properties of the field; we will call it the Lagrangian potential.

Varying $\underline{\phi}$ and $\overline{\lambda}$ as independent variables in the action (2.9), we get the field equations

$$\nabla\underline{\phi} = \frac{\partial \Lambda}{\partial \overline{\lambda}}, \qquad \nabla \circ \overline{\lambda} = -\frac{\partial \Lambda}{\partial \underline{\phi}}, \tag{2.10}$$

or, writing the same equations in the coordinate basis,

$$\phi_{\{\alpha\};\beta} = \frac{\partial \Lambda}{\partial \lambda^{[\alpha]\beta}}, \qquad \lambda^{\{\alpha\}\beta}{}_{;\beta} = -\frac{\partial \Lambda}{\partial \phi_{\{\alpha\}}}. \tag{2.11}$$

Eliminating $\overline{\lambda}$ from these equations and using the property

$$-\frac{\partial \Lambda(\underline{\phi}, \overline{\lambda})}{\partial \underline{\phi}} = \frac{\partial L(\underline{\phi}, \nabla\underline{\phi})}{\partial \underline{\phi}} \tag{2.12}$$

of the Legendre dual transformation, we return to the second-order field equations (2.4).

Sometimes, however, we cannot calculate all $\phi_{\{\alpha\};\beta}$ in terms of $\lambda_{\{\alpha\}\beta}$ from Eq. (2.5). This happens when L does not depend on all derivatives $\phi_{\{\alpha\};\beta}$, but only on some combination of them. As an important and typical example, take a covector field ϕ_α and decompose its covariant derivatives $\phi_{\alpha;\beta}$ into symmetrical and antisymmetrical parts,

$$\begin{aligned} &\phi_{\alpha;\beta} = \phi_{(\alpha;\beta)} + \phi_{[\alpha;\beta]}, \\ &\phi_{(\alpha;\beta)} = \tfrac{1}{2}(\phi_{\alpha;\beta} + \phi_{\beta;\alpha}), \quad \phi_{[\alpha;\beta]} = \tfrac{1}{2}(\phi_{\alpha;\beta} - \phi_{\beta;\alpha}). \end{aligned} \tag{2.13}$$

Let now L depend only on the antisymmetrical combination $\phi_{[\alpha;\beta]}$. From Eq. (2.5), we see that the symmetrical part of $\lambda^{\alpha\beta}$ automatically vanishes, $\lambda^{(\alpha\beta)} = 0$. Therefore, we cannot calculate all $\phi_{\alpha;\beta}$ from Eq. (2.5). However, we can calculate exactly that combination of $\phi_{\alpha;\beta}$ which we need, namely $\phi_{[\alpha;\beta]}$, by inverting the equation

$$\lambda^{[\alpha\beta]} = \frac{\partial L}{\partial \phi_{[\alpha;\beta]}}. \tag{2.14}$$

We then perform the Legendre transformation in the variables $\phi_{[\alpha;\beta]}$ only, getting

$$\begin{aligned} &\Lambda \equiv \lambda^{[\alpha\beta]} \phi_{[\alpha;\beta]} - L(\phi_\alpha, \phi_{[\alpha;\beta]}, g_{\alpha\beta}), \\ &\phi_{[\alpha;\beta]} = \phi_{[\alpha;\beta]}(\phi_\gamma, \lambda^{[\gamma\delta]}, g_{\gamma\delta}). \end{aligned} \tag{2.15}$$

We obtain the field equations by varying the first-order action

$$S^{\Phi}=S^{\Phi}[\phi_{\alpha},\ \lambda^{[\alpha\beta]}]=\int_{M}\underline{\eta}(\lambda^{[\alpha\beta]}\phi_{\alpha;\beta}-\Lambda) \tag{2.16}$$

with respect to ϕ_{α} and *antisymmetrical* $\lambda^{[\alpha\beta]}$ as independent variables.

A similar procedure is applicable whenever L depends on an irreducible combination of the covariant derivatives $\phi_{\{\alpha\};\beta}$ rather than on all such derivatives. With these remarks in mind, we return to the treatment of the general case in which Eq. (2.5) is fully invertible.

3. SYMMETRICAL AND CANONICAL ENERGY-MOMENTUM TENSORS

Long time ago, Belinfante,[5] Rosenfeld,[6] and Pauli[7] discovered the connection between the symmetrical energy–momentum tensor $T^{\alpha\beta}$, which is the source of the gravitational field $g_{\alpha\beta}$, and the canonical energy-momentum tensor $\Theta^{\alpha\beta}$, which follows from the field Lagrangian by standard procedures of the Hamiltonian theory. We shall review the basic results of the Belinfante–Rosenfeld approach adopted to the first-order action, because they are essential for the proper understanding of the structure of the field super-Hamiltonian H^{Φ} and supermomentum $\tilde{H}^{\Phi}$ in hypersurface dynamics.

The symmetrical energy-momentum tensor is defined as the variational derivative of the field action with respect to the spacetime metric,

$$|{}^4g|^{1/2}T^{\alpha\beta}(X)\equiv 2\,\frac{\delta S^{\Phi}}{\delta g_{\alpha\beta}(X)}\ , \tag{3.1}$$

the factor 2 being inserted to get the correct coefficients in the Einstein's law (our units are $2\kappa\equiv 16\pi G\,c^{-4}=1$; see the final paper of this series). It is well-known that the energy–momentum tensor (3.1) is covariantly conserved by virtue of the field equations (2.4) or (2.11),

$$\nabla\cdot\overline{T}=0,\quad T^{\alpha\beta}{}_{;\beta}=0, \tag{3.2}$$

because the field action (2.1) or (2.9) is left invariant by spacetime diffeomorphisms.[8]

To find the detailed structure of the symmetrical energy–momentum tensor, vary the first-order action (2.9) with respect to the metric tensor $\underline{g}$,

$$\delta_g S^{\Phi}=\int_{M}\underline{\eta}\left\{\left[-\frac{\partial\Lambda}{\partial g_{\mu\nu}}+\tfrac{1}{2}(\lambda^{\{\alpha\}\beta}\phi_{\{\alpha\};\beta}-\Lambda)\,g^{\mu\nu}\right]\delta g_{\mu\nu}+\lambda^{\{\alpha\}\lambda}\,\delta_g\phi_{\{\alpha\};\lambda}\right\}. \tag{3.3}$$

Here, the second term in the [ ] brackets comes from the variation of the Levi-Civita form $\underline{\eta}$. To evaluate the variation $\delta_g\phi_{\{\alpha\};\lambda}$, write the covariant derivative $\phi_{\{\alpha\};\lambda}$ as

$$\phi_{\{\alpha\};\lambda}=\phi_{\{\alpha\},\lambda}-\mathcal{T}_{\{\alpha\}}{}^{\{\beta\}}{}_{\iota}{}^{\kappa}\,\phi_{\{\beta\}}\Gamma^{\iota}{}_{\kappa\lambda}, \tag{3.4}$$

where $\mathcal{T}_{\{\alpha\}}{}^{\{\beta\}}{}_{\iota}{}^{\kappa}$ is a combination of Kronecker's deltas,

$$\mathcal{T}_{\{\alpha\}}{}^{\{\beta\}}{}_{\iota}{}^{\kappa}=\sum_{r=1}^{n}\delta^{\beta_1}_{\alpha_1}\cdots\delta^{\beta_{r-1}}_{\alpha_{r-1}}\delta^{\beta_{r+1}}_{\alpha_{r+1}}\cdots\delta^{\kappa}_{\alpha_r}\delta^{\beta_r}_{\iota}. \tag{3.5}$$

The variation of $\phi_{\{\alpha\};\lambda}$ is then reduced to the variation of the affine connection $\Gamma^{\iota}{}_{\kappa\lambda}$,

$$\delta_g\phi_{\{\alpha\};\lambda}=-\mathcal{T}_{\{\alpha\}}{}^{\{\beta\}}{}_{\iota}{}^{\kappa}\phi_{\{\beta\}}\,\delta_g\Gamma^{\iota}{}_{\kappa\lambda}, \tag{3.6}$$

which is given by the well-known formula

$$\delta_g\Gamma^{\iota}{}_{\kappa\lambda}=\tfrac{1}{2}(-g^{\iota\sigma}\delta^{\mu\nu}_{\kappa\lambda}+g^{\iota\mu}\delta^{\nu\sigma}_{\kappa\lambda}+g^{\iota\nu}\delta^{\mu\sigma}_{\kappa\lambda})\cdot(\delta g_{\mu\nu})_{;\sigma}, \tag{3.7}$$

$\delta^{\mu\nu}_{\kappa\lambda}\equiv\delta^{(\mu}_{\kappa}\delta^{\nu)}_{\lambda}$.

From Eqs. (3.6) and (3.7), we get

$$\lambda^{\{\alpha\}\lambda}\delta_g\phi_{\{\alpha\};\lambda}=\tfrac{1}{2}P^{\mu\nu\sigma}\delta g_{\mu\nu;\sigma}, \tag{3.8}$$

where

$$P^{\mu\nu\sigma}\equiv(-\lambda^{\{\alpha\}\sigma}\mathcal{T}_{\{\alpha\}}{}^{\{\beta\}\,(\mu\nu)}+\lambda^{\{\alpha\}\mu}\mathcal{T}_{\{\alpha\}}{}^{\{\beta\}\,[\sigma\nu]}+\lambda^{\{\alpha\}\nu}\mathcal{T}_{\{\alpha\}}{}^{\{\beta\}\,[\sigma\mu]})\,\phi_{\{\beta\}} \tag{3.9}$$

is a bilinear form of $\lambda^{\{\alpha\}\lambda}$ and $\phi_{\{\beta\}}$, symmetrical in the indices $\mu\nu$. Integrating the term (3.8) in Eq. (3.3) by parts, we identify the symmetrical energy–momentum tensor (3.1) as

$$T^{\mu\nu}=-2\,\frac{\partial\Lambda}{\partial g_{\mu\nu}}+(\lambda^{\{\alpha\}\beta}\phi_{\{\alpha\};\beta}-\Lambda)\,g^{\mu\nu}-P^{\mu\nu\sigma}{}_{;\sigma}. \tag{3.10}$$

To find the relation of the symmetrical energy–momentum tensor (3.10) to the canonical energy–momentum tensor $\Theta^{\alpha\beta}$, write down the condition ensuring that Λ behaves as a scalar under the change

$$\overline{\partial}_{\alpha}\to\overline{e}_{\alpha}=A^{\beta'}_{\alpha}\overline{\partial}_{\beta'} \tag{3.11}$$

of the spacetime basis at a given point X. During the transformation (3.11), the components $\phi_{\{\alpha\}}$ and $\phi^{\{\alpha\}}$ of covariant and contravariant tensors, respectively, undergo the change

$$\phi_{\{\alpha'\}}=A^{\{\beta\}}_{\{\alpha'\}}\,\phi_{\{\beta\}},\quad A^{\{\beta\}}_{\{\alpha'\}}\equiv A^{\beta_1}_{\alpha'_1}\cdots A^{\beta_n}_{\alpha'_n}, \tag{3.12}$$

and

$$\phi^{\{\alpha'\}}=A^{\{\alpha'\}}_{\{\beta\}}\phi^{\{\beta\}},\quad A^{\{\alpha'\}}_{\{\beta\}}\equiv A^{\alpha'_1}_{\beta_1}\cdots A^{\alpha'_n}_{\beta_n}, \tag{3.13}$$

where $A^{\alpha'}_{\beta}$ is the inverse matrix to $A^{\alpha}_{\beta'}$,

$$A^{\alpha'}_{\beta}A^{\beta}_{\gamma'}=\delta^{\alpha'}_{\gamma'}. \tag{3.14}$$

Differentiating Eqs. (3.12)–(3.14) with respect to $A^{\mu'}_{\nu}$, we get

$$\left.\frac{\partial\phi_{\{\alpha'\}}}{\partial A^{\mu'}_{\nu}}\right|_{A^{\mu'}_{\nu}=\delta^{\mu}_{\nu}}=\mathcal{T}_{\{\alpha\}}{}^{\{\beta\}}{}_{\mu}{}^{\nu}\,\phi_{\{\beta\}}, \tag{3.15}$$

and

$$\left.\frac{\partial\phi^{\{\alpha'\}}}{\partial A^{\mu'}_{\nu}}\right|_{A^{\mu'}_{\nu}=\delta^{\mu}_{\nu}}=-\mathcal{T}_{\{\beta\}}{}^{\{\alpha\}}{}_{\mu}{}^{\nu}\,\phi^{\{\beta\}}. \tag{3.16}$$

The expression $\Lambda(\phi_{\{\alpha\}},\lambda^{\{\alpha\}\beta},g_{\alpha\beta})$ behaves as a scalar iff

$$\left.\frac{\partial\Lambda}{\partial A^{\mu'}_{\nu}}\right|_{A^{\mu'}_{\nu}=\delta^{\mu}_{\nu}}=\left[\frac{\partial\Lambda}{\partial\phi_{\{\alpha'\}}}\frac{\partial\phi_{\{\alpha'\}}}{\partial A^{\mu'}_{\nu}}+\frac{\partial\Lambda}{\partial\lambda^{\{\alpha'\}\beta'}}\frac{\partial\lambda^{\{\alpha'\}\beta'}}{\partial A^{\mu'}_{\nu}}+\frac{\partial\Lambda}{\partial g_{\alpha'\beta'}}\frac{\partial g_{\alpha'\beta'}}{\partial A^{\mu'}_{\nu}}\right]_{A^{\mu'}_{\nu}=\delta^{\mu}_{\nu}}=0. \tag{3.17}$$

Substituting here the appropriate expressions from Eqs. (3.15), (3.16), we can express $\partial\Lambda/\partial g_{\mu\nu}$ as

$$2\,\frac{\partial\Lambda}{\partial g_{\mu\nu}}\equiv-\frac{\partial\Lambda}{\partial\phi_{\{\alpha\}}}\mathcal{T}_{\{\alpha\}}{}^{\{\beta\}\mu\nu}\phi_{\{\beta\}}+\frac{\partial\Lambda}{\partial\lambda^{\{\alpha\}\gamma}}\mathcal{T}_{\{\beta\}}{}^{\{\alpha\}\mu\nu}\lambda^{\{\beta\}\gamma}+g^{\mu\sigma}\frac{\partial\Lambda}{\partial\lambda^{\{\alpha\}\sigma}}\lambda^{\{\alpha\}\nu}. \tag{3.18}$$

Using the field equations (2.11) in the identity (3.18), we cast it into the form

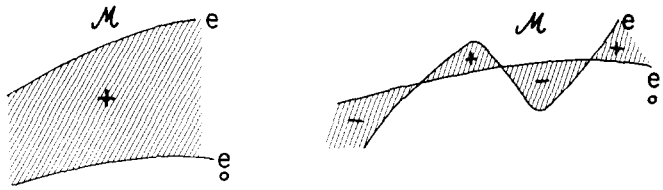

FIG. 1a. Sandwich action. The sandwich action $S^{\Phi}[\underset{\circ}{e}, e]$ is defined as the field action contained in the spacetime sandwich enclosed between two embeddings, $\underset{\circ}{e}$ and e. The contributions are taken with the positive sign in the regions where e lies to the future of $\underset{\circ}{e}$ and with the negative sign in the regions where e lies to the past of $\underset{\circ}{e}$.

$$2\frac{\partial\Lambda}{\partial g_{\mu\nu}} = \phi_{\{\alpha\}}{}^{;\mu}\lambda^{\{\alpha\}\nu} + \phi_{\{\alpha\};\gamma}\mathcal{T}_{\{\beta\}}{}^{\{\alpha\}\mu\nu}\lambda^{\{\beta\}\gamma} + \lambda^{\{\alpha\}\gamma}{}_{;\gamma}\mathcal{T}_{\{\alpha\}}{}^{\{\beta\}\mu\nu}\phi_{\{\beta\}}. \tag{3.19}$$

Substituting the expression (3.19) into Eq. (3.10), we express $T^{\mu\nu}$ in the form

$$T^{\mu\nu} = \Theta^{\mu\nu} + S^{\mu\nu\sigma}{}_{;\sigma}, \tag{3.20}$$

where

$$\Theta_{\mu}{}^{\nu} = -\phi_{\{\alpha\};\mu}\lambda^{\{\alpha\}\nu} + (\lambda^{\{\alpha\}\beta}\phi_{\{\alpha\};\beta} - \Lambda)\,\delta^{\nu}_{\mu} \tag{3.21}$$

is known as the canonical energy–momentum tensor, and

$$S^{\mu\nu\sigma} = -P^{\mu\nu\sigma} - \lambda^{\{\alpha\}\sigma}\mathcal{T}_{\{\alpha\}}{}^{\{\beta\}\mu\nu}\phi_{\{\beta\}} = (-\lambda^{\{\alpha\}\sigma}\mathcal{T}_{\{\alpha\}}{}^{\{\beta\}[\mu\nu]} + \lambda^{\{\alpha\}\mu}\mathcal{T}_{\{\alpha\}}{}^{\{\beta\}[\nu\sigma]} + \lambda^{\{\alpha\}\nu}\mathcal{T}_{\{\alpha\}}{}^{\{\beta\}[\mu\sigma]})\,\phi_{\{\beta\}} \tag{3.22}$$

is a bilinear form of $\lambda^{\{\alpha\}\beta}$ and $\phi_{\{\alpha\}}$ antisymmetrical in the indices $\nu\sigma$, called the spin tensor. Its divergence, $S^{\mu\nu\sigma}{}_{;\sigma}$, is called the spin energy–momentum tensor.

The canonical energy–momentum tensor is not covariantly conserved in curved spacetimes. Indeed, from the conservation law (3.2) and the commutation relations (II. 6. 1) (see Ref. 2) for the covariant derivatives, we get

$$\Theta_{\mu}{}^{\nu}{}_{;\nu} = -S_{\mu}{}^{\nu\sigma}{}_{;\sigma\nu} = S_{\mu}{}^{\nu\sigma}{}_{;[\nu\sigma]} = -\tfrac{1}{2}R_{\mu\nu\sigma\tau}S^{\nu\sigma\tau}. \tag{3.23}$$

The projections of the symmetrical and spin energy–momentum tensors play an important role in hypersurface dynamics, where they are identified with the different pieces of the super-Hamiltonian and supermomentum. We shall return to this subject in Sec. 8.

4. HYPERSURFACE ACTION

The action (2.9) is a functional of the spacetime tensor fields $\underset{\circ}{\phi}(X) \in \mathcal{T}^0_n(\mathcal{M})$, $\underset{\circ}{\lambda}(X) \in \mathcal{T}^{n+1}_0(\mathcal{M})$ and $\underset{\circ}{g}(X) \in \mathcal{T}^0_2(\mathcal{M})$. We shall now consider the action S^{Φ} also as a functional of the embedding. This is done by limiting the region of integration $\mathcal{M}$ to a sandwich of spacetime enclosed between two embeddings, $\underset{\circ}{e}$ and e. We adopt the convention that this integral is taken with a positive sign when e lies to the future of $\underset{\circ}{e}$, and with a negative sign when e lies to the past of $\underset{\circ}{e}$. If e and $\underset{\circ}{e}$ intersect, we have alternating regions in which the integral is taken with positive and negative signs [Fig. 1(a)]. We write

$$S^{\Phi}[\underset{\circ}{e}, e] = \int_{\underset{\circ}{e}}^{e} \underset{\sim}{\eta} L \tag{4.1}$$

for the action enclosed between the embeddings $\underset{\circ}{e}$ and e. In the limit when the initial embedding is pushed far back into the past, we write

$$S^{\Phi}[e] = \int_{e_{-\infty}}^{e} \underset{\sim}{\eta} L, \tag{4.2}$$

the integration being performed over the entire past of the embedding e.

The action (4.2) is a functional of the embedding, $S^{\Phi}[e] \in \mathcal{F}(\mathcal{E})$. It is obvious, however, that if we change the embedding e while leaving the hypersurface fixed, $e \to e \circ \varphi$, $\varphi \in \mathrm{Diff}(m)$, the action $S^{\Phi}[e]$ remains the same. The action (4.2) may thus be considered as a functional $S^{\Phi}[h] \in \mathcal{F}(\mathcal{H})$ of the hypersurface, and we can write

$$S^{\Phi}[h] = \int_{h_{-\infty}}^{h} \underset{\sim}{\eta} L.$$

The same remark applies to the action (4.1). In the differential language (Sec. I. 4) (see Ref. 1) we have

$$\delta_{\tilde{N}} S^{\Phi}[e] = 0 \quad \forall \tilde{N}. \tag{4.3}$$

Let us adopt the convention that the normal $\bar{n}$ to the embedding e points into the future and the embedding is oriented so that the vectors $\{\bar{n}, \bar{e}_1, \bar{e}_2, \bar{e}_3\}$ form a right-handed system. Projecting the Levi-Civita form $\underset{\sim}{\eta}$, we get

$$\eta_{\perp bcd} \equiv \epsilon\eta_{\alpha\beta\gamma\delta}\, n^{\alpha} e^{\beta}_{b} e^{\gamma}_{c} e^{\delta}_{d} = \epsilon\eta_{bcd}, \tag{4.4}$$

and, of course,

$$\eta_{abcd} = 0. \tag{4.5}$$

Therefore,

$$\eta_{\alpha bcd} = n_{\alpha}\eta_{\perp bcd} = \epsilon n_{\alpha}\eta_{bcd}. \tag{4.6}$$

Asking how the action (4.2) changes under the deformation $\mathbf{N}$ of the embedding e, we get

$$\delta_{\mathbf{N}} S^{\Phi}[e] = \langle \mathbf{L}, \mathbf{N} \rangle = \int_{m} N^{\alpha}\eta_{\alpha bcd} L\, \underset{\sim}{dx}{}^{a} \wedge \underset{\sim}{dx}{}^{b} \wedge \underset{\sim}{dx}{}^{c} = \delta_{N} S^{\Phi}[e] = \int_{m} \underset{\sim}{\eta} N L. \tag{4.7}$$

Here, $\langle \mathbf{L}, \mathbf{N} \rangle$ is the inner product in $\mathcal{E}$ between the deformation $\mathcal{E}$-vector $\mathbf{N}$ and the $\mathcal{E}$-covector

$$\mathbf{L} = \mathrm{d}S^{\Phi}[e] \tag{4.8}$$

with the coordinate $\mathcal{E}$-basis components

$$L_{\alpha bcd} = \eta_{\alpha bcd} L. \tag{4.9}$$

To get the final line of Eq. (4.7), we have used Eq.

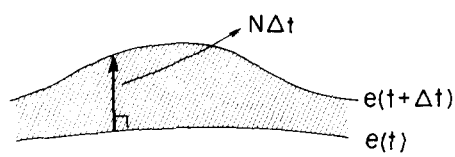

FIG. 1b. Hypersurface Lagrangian. The hypersurface Lagrangian $\delta_N S^{\Phi}$ is interpreted as $(\Delta t)^{-1}$ times the action contained in the thin sandwich enclosed by the nearby embeddings $e(t)$ and $e(t + \Delta t)$ with the normal proper time separation $N\Delta t$.

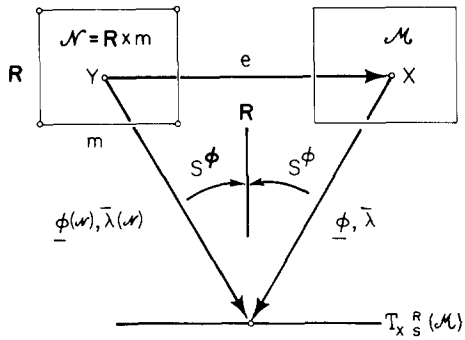

FIG. 2. Hypersurface action. The commutative diagram defining the hypersurface action S^Φ in terms of the spacetime action S^ϕ.

(4.6), checking thus directly that $S^\phi[e]$ is an $\mathcal{H}$-scalar, Eq. (4.3).

We shall call $\delta_{\mathbf{N}} S^\phi[e]$ the hypersurface Lagrangian and $\mathbf{L}$ the Lagrangian $\mathcal{H}$-covector. The expression $\Delta t\ \delta_{\mathbf{N}} S^\phi[e]$ can be interpreted as the field action contained in a thin spacetime sandwich enclosed between the hypersurfaces $\{e(t)\}$ and $\{e(t+\Delta t)\}$ which have the normal proper time separation $N\Delta t$ [Fig. 1(b)].

To recover the field action contained in a finite sandwich between the hypersurfaces $\{\underset{1}{e}\}$ and $\{\underset{2}{e}\}$, we connect $\underset{1}{e}$ and $\underset{2}{e}$ by a path $e(t)$ in $\mathcal{E}$ and integrate the hypersurface Lagrangian along this path,

$$S^\Phi = \int_{\underset{1}{t}}^{\underset{2}{t}} \langle \mathbf{L}, \mathbf{N} \rangle\, dt. \tag{4.10}$$

Consider the path $e(t)$ as a mapping e from the manifold $\mathcal{N} \equiv \mathbf{R}\times m$ into the spacetime $\mathcal{M}$,

$$e : Y=(t,x)\in\mathcal{N} \to X = e(t,x)\in\mathcal{M}. \tag{4.11}$$

For the rest of this section, e will not denote an individual embedding, but the whole path (4.11), the individual embedding corresponding to a fixed t being denoted by e_t. For simplicity, we shall assume that $e(t)$ is a foliation of $\mathcal{M}$, i.e., that the mapping e is a diffeomorphism. In the final results, we can waive this simplifying assumption and consider again any path $e(t)$ in $\mathcal{E}$ (this will be done in Sec. 6).

The mapping e refers the fields $\underline{\phi}(X)$ and $\bar{\chi}(X)$ back to the manifold $\mathcal{N}$,

$$\underline{\phi}(Y) = \underline{\phi}(e(Y)),\quad \bar{\chi}(Y) = \bar{\chi}(e(Y)). \tag{4.12}$$

Note that the fields $\underline{\phi}(Y)$ and $\bar{\chi}(Y)$ defined by Eq. (4.12) are considered as the fields of *spacetime* tensors, and are thus different from the fields $e^*\underline{\phi}$ and $e^{-1}{}_*\,\bar{\chi}$, which are tensor fields on $\mathcal{N}$. For a fixed t, Eqs. (4.12) define the spacetime hypertensors ϕ_t and λ_t along the embedding e_t. We can thus think about the fields (4.12) as hypertensors specified along the path $e(t)$. In this spirit, we write

$$\begin{aligned} &\phi : Y\in\mathcal{N} \to \underline{\phi}(Y)\in T_{e(Y)}{}^0_n(\mathcal{M}),\\ &\lambda : Y\in\mathcal{N} \to \bar{\chi}(Y)\in T_{e(Y)}{}^{n+1}_0(\mathcal{M}). \end{aligned} \tag{4.13}$$

The action functional S^Φ maps the fields $\underline{\phi}(X)$, $\bar{\chi}(X)$ on $(\mathcal{M}, \underline{g})$ into real numbers. For a given path e, we define the functional S^Φ so that it assigns the same number to the fields ϕ, λ as the action functional S^Φ did to the fields $\underline{\phi}$, $\bar{\chi}$:

$$S^\Phi[\phi, \lambda; e] = S^\Phi[\underline{\phi}, \bar{\chi}; \underline{g}]. \tag{4.14}$$

This is expressed by the commutative diagram in Fig. 2. Because we have specialized e to a diffeomorphism, S^Φ is uniquely defined for every field ϕ, λ. Applying the rule (4.14) to the action (4.1), we get exactly the action (4.10). We shall call the action $S^\Phi[\phi, \lambda, e]$ expressed as a functional of the hypertensors ϕ, λ along an embedding e the *hypersurface action.*

From Eq. (4.14), we immediately see that $\phi(Y)$, $\lambda(Y)$ extremizes the functional S^Φ evaluated along a fixed path e if and only if $\underline{\phi}(X)$, $\bar{\chi}(X)$ extremizes the functional S^ϕ on a fixed Riemannian background $(\mathcal{M}, \underline{g})$. Moreover, S^Φ is obviously unchanged if we keep the spacetime fields $\underline{\phi}(X)$, $\bar{\chi}(X)$, and $\underline{g}(X)$ fixed, but change the path e between the fixed end points $\underset{1}{e}, \underset{2}{e}$. Such a change, of course, induces the change in the fields $\phi(Y)$, $\lambda(Y)$,

$$\begin{aligned} &\delta_e \phi_{\{\alpha\}}(Y) = \phi_{\{\alpha\},\gamma}(X)\big|_{X=e(Y)}\, \delta e^\gamma(Y),\\ &\delta_e \lambda^{\{\alpha\}\beta}(Y) = \lambda^{\{\alpha\}\beta}{}_{,\gamma}(X)\big|_{X=e(Y)}\, \delta e^\gamma(Y). \end{aligned} \tag{4.15}$$

Under the variation (4.15), the functional S^Φ is left unchanged for any ϕ, λ, and e:

$$\frac{\delta S}{\delta \phi^Y}\,\delta_e \phi^Y + \frac{\delta S}{\delta \lambda^Y}\,\delta_e \lambda^Y + \frac{\delta S}{\delta e^Y}\,\delta e^Y \equiv 0. \tag{4.16}$$

If $\phi(Y)$, $\lambda(Y)$ extremize the functional S^Φ along a given path e, the first two variational derivatives in Eq. (4.16) must vanish. We thus conclude that

$$\frac{\delta S^\Phi}{\delta e(Y)} = 0 \tag{4.17}$$

for any $e(Y)$ and the extremal fields $\phi(Y)$, $\lambda(Y)$. Equation (4.17) is a consequence of the field equations, due to the identity (4.16). This means we get the correct equations by varying the action functional $S^\Phi[\phi, \lambda; e]$ with respect to all the variables ϕ, λ, e. These equations, however, are not all independent, being connected by the identity (4.16). The path $e(Y)$ cannot be determined from them, but may be prescribed arbitrarily. The equations then determine the extremal field $\phi(Y)$, $\lambda(Y)$ along the given path $e(Y)$.

At this point, we can project the fields $\underline{\phi}(Y)$ and $\bar{\chi}(Y)$ into $\perp$ and $\parallel$ directions to the embeddings e_t, i.e., we can split the hypertensors $\phi(x)[e_t]$ and $\lambda(x)[e_t]$ with respect to the normal hyperbasis. We can then vary the projections $\phi_{\perp,\parallel}(Y)$, $\lambda^{\perp,\parallel}(Y)$ [keeping $e(Y)$ fixed or varying it as well], instead of varying the original variables $\phi(Y)$, $\lambda(Y)$ [keeping $e(Y)$ fixed or varying it as well]. Because $\{\bar{n}, \bar{e}\}$ are some functionals of $e(Y)$ in a given Riemannian spacetime $(\mathcal{M}, \underline{g})$, the new variables are some functionals of the old variables. Moreover, the process can be easily inverted, and the old variables expressed as some functionals of the new variables.

The equations which we get by varying S^Φ with respect to the projected variables $\phi_{\perp,\parallel}(Y)$ and $\lambda^{\perp,\parallel}(Y)$ with $e(Y)$ kept fixed are thus equivalent to the original

field equations. Also, we may decide to vary $e(Y)$ in addition to $\phi_{\perp,\parallel}(Y)$ and $\lambda^{\perp,\parallel}(Y)$, and still get a correct (though redundant) set of equations. These are the main conclusions of this section.

Our basic technical task then is to express the hypersurface Lagrangian in terms of the projections $\phi_{\perp,\parallel}(Y)$, $\lambda^{\perp,\parallel}(Y)$ and the path $e(Y)$, and cast it into the Hamiltonian form. This is done in the next section.

5. HYPERSURFACE LAGRANGIAN OF A COVECTOR FIELD

The simplest field on which the projection of the hypersurface Lagrangian may be illustrated with all its complexities is a covector field $\underline{\phi}(X) \in \mathcal{T}_1^0(\mathcal{M})$. The scalar field $\phi(X) \in \mathcal{T}(\mathcal{M})$ is too special for this purpose, because the covariant derivatives $\nabla\phi$ may be replaced by the exterior derivatives $\underline{d}\phi$. The scalar field action thus does not depend on the derivatives of the metric tensor; we say that the scalar field has a nonderivative gravitational coupling.[9] We will treat it, together with other fields sharing this property, in Sec. 11.

The hypersurface Lagrangian of a covector field has the form

$$\delta_N S^\phi = \int_m \underset{\sim}{\eta}\, N(\bar{\lambda} \rfloor \nabla \underline{\phi} - \Lambda(\underline{\phi}, \bar{\lambda}, \underline{g})). \tag{5.1}$$

We will rearrange it into the Hamiltonian form in four steps. In the first step, we project the spacetime fields $\underline{\phi}(X)$, $\bar{\lambda}(X)$, and $\underline{g}(X)$, substitute the projections into the Lagrangian potential Λ, and write down the conditions that Λ expressed in this way is a spacetime scalar. In the second step, we project the bilinear form $Ng^{1/2}\lambda^{\alpha\beta}\phi_{\alpha;\beta}$. In the third step, we identify the momenta conjugate to the projections $\phi_\perp$ and ϕ_a as the coefficients of $\delta_{\mathbf{N}}\phi_\perp$ and $\delta_{\mathbf{N}}\phi_a$ in the hypersurface Lagrangian. In the fourth step, we integrate by parts those terms which contain the space derivatives of the lapse function and the shift vector, and identify the field super-Hamiltonian and supermomentum.

Start with the first step of this program. The fields $\underline{\phi}(X)$, $\bar{\lambda}(X)$, and $\underline{g}(X)$ in Eq. (5.1) are spacetime tensor fields defined along the embedding $X = e(x)$; according to the terminology introduced in Sec. I.7, each of these fields is a hypertensor. We split these hypertensors with respect to the normal hyperbasis $\{\mathbf{n}, \mathbf{e}\}$, i.e., decompose the constituent spacetime tensors with respect to the normal basis $\{\vec{n}, \vec{e}\}$ according to the scheme of Sec. I.3,

$$\begin{aligned}
&\phi_\alpha = \phi_\perp n_\alpha + \phi_a e^a_\alpha,\\
&\lambda^{\alpha\beta} = \lambda^{\perp\perp} n^\alpha n^\beta + \lambda^{a\perp} e^\alpha_a n^\beta + \lambda^{\perp b} n^\alpha e^\beta_b + \lambda^{ab} e^\alpha_a e^\beta_b,\\
&g_{\alpha\beta} = g_{ab} e^a_\alpha e^b_\beta + \epsilon n_\alpha n_\beta.
\end{aligned} \tag{5.2}$$

The Lagrangian potential Λ may be expressed in terms of these projections,

$$\Lambda = \Lambda(\phi_\perp, \phi_a, \lambda^{\perp\perp}, \lambda^{\perp b}, \lambda^{a\perp}, \lambda^{ab}). \tag{5.3}$$

Because Λ is a spacetime scalar, the function (5.3) must be a space scalar, and an invariant under the hypersurface tilts (see Sec. II.3). The first condition leads to an identity of the type (3.18), with the space indices replacing the spacetime indices,

$$\begin{aligned}
2\frac{\partial\Lambda}{\partial g_{mn}} = &-\frac{\partial\Lambda}{\partial\phi_n}\phi^m + \frac{\partial\Lambda}{\partial\lambda^{a\perp}} g^{am}\lambda^{n\perp} + \frac{\partial\Lambda}{\partial\lambda^{\perp a}} g^{am}\lambda^{\perp n}\\
&+ \frac{\partial\Lambda}{\partial\lambda^{ab}} g^{bm}\lambda^{an} + \frac{\partial\Lambda}{\partial\lambda^{ab}} g^{am}\lambda^{nb}.
\end{aligned} \tag{5.4}$$

The second condition states that

$$\begin{aligned}
\delta_\mathcal{F}\Lambda = &\frac{\partial\Lambda}{\partial\phi_\perp}\delta_\mathcal{F}\phi_\perp + \frac{\partial\Lambda}{\partial\phi_a}\delta_\mathcal{F}\phi_a + \frac{\partial\Lambda}{\partial\lambda^{\perp\perp}}\delta_\mathcal{F}\lambda^{\perp\perp} + \frac{\partial\Lambda}{\partial\lambda^{\perp b}}\delta_\mathcal{F}\lambda^{\perp b}\\
&+ \frac{\partial\Lambda}{\partial\lambda^{a\perp}}\delta_\mathcal{F}\lambda^{a\perp} + \frac{\partial\Lambda}{\partial\lambda^{ab}}\delta_\mathcal{F}\lambda^{ab} = 0
\end{aligned} \tag{5.5}$$

for an arbitrary hypersurface tilt $N(x) = 0$, $N_{,c}(x) \neq 0$. Using Eqs. (II.3.1) and (II.3.16) for the tilt changes of tensor projections, and taking into account [as we already did when writing Eq. (5.5)] that $\delta_\mathcal{F} g_{ab} = 0$, we get the identity

$$\begin{aligned}
&-\frac{\partial\Lambda}{\partial\phi_\perp}\phi^c + \epsilon\frac{\partial\Lambda}{\partial\phi_c}\phi_\perp - \frac{\partial\Lambda}{\partial\lambda^{\perp\perp}}(\lambda^{\perp c} + \lambda^{c\perp})\\
&+ \frac{\partial\Lambda}{\partial\lambda^{\perp b}}(\epsilon\lambda^{\perp\perp} g^{bc} - \lambda^{cb}) + \frac{\partial\Lambda}{\partial\lambda^{a\perp}}(\epsilon\lambda^{\perp\perp} g^{ac} - \lambda^{ac})\\
&+ \epsilon\frac{\partial\Lambda}{\partial\lambda^{ab}}(\lambda^{a\perp} g^{bc} + \lambda^{\perp b} g^{ac}) = 0.
\end{aligned} \tag{5.6}$$

The identities (5.4) and (5.6) may also be considered as the $\parallel\parallel$ and $\parallel\perp$ projections of the spacetime identity (3.18).

The second step of our program is to project the bilinear form $N\lambda^{\alpha\beta}\phi_{\alpha;\beta}$. We write

$$\begin{aligned}
N\lambda^{\alpha\beta}\phi_{\alpha;\beta} = &\lambda^{\perp\perp}(N\phi_{\perp;\perp}) + \epsilon\lambda^{a\perp}(N\phi_{a;\perp})\\
&+ \epsilon N\lambda^{\perp a}\phi_{\perp;a} + N\lambda^{ab}\phi_{a;b}
\end{aligned} \tag{5.7}$$

and use Eqs. (II.2.5) and (II.2.6) for the projections of the covariant derivatives. This yields

$$\begin{aligned}
N\underset{\sim}{\eta}\lambda^{\alpha\beta}\phi_{\alpha;\beta} = &\epsilon\underset{\sim}{\eta}\lambda^{\perp\perp}\delta_N\phi_\perp + \underset{\sim}{\eta}\lambda^{a\perp}\delta_N\phi_a\\
&+ \underset{\sim}{\eta}(\lambda^{ab}\phi_{a|b} + \epsilon\lambda^{\perp a}\phi_{\perp|a})N\\
&- 2K_{ab}\underset{\sim}{P}^{ab}N\\
&+ \epsilon\underset{\sim}{\eta}(\lambda^{\perp\perp}\phi^a - \lambda^{a\perp}\phi_\perp)N_{,a}.
\end{aligned} \tag{5.8}$$

We have introduced the symmetrical bilinear form $\underset{\sim}{P}^{ab}$ in the projections of the $\bar{\lambda}$ and $\underline{\phi}$ tensors by the formula

$$\underset{\sim}{P}^{ab} = \tfrac{1}{2}\underset{\sim}{\eta}(-\lambda^{\perp(a}\phi^{b)} + \lambda^{(ab)}\phi_\perp - \lambda^{(a\perp}\phi^{b)}). \tag{5.9}$$

In the third step, we replace the normal changes δ_N of the projections $\phi_\perp$ and ϕ_a by the changes $\delta_{\mathbf{N}}$ along an arbitrary deformation $\mathcal{E}$-vector $\mathbf{N}$, and use the fact that the tangential changes $\delta_{\tilde{N}}$ of the hypervectors $\phi_\perp$ and ϕ_a are equal to the Lie derivatives $L_{\tilde{N}}$ [cf. Eqs. (II.5.3) and (II.5.4)],

$$\begin{aligned}
&\delta_N\phi_\perp = \delta_{\mathbf{N}}\phi_\perp - \delta_{\tilde{N}}\phi_\perp = \delta_{\mathbf{N}}\phi_\perp - L_{\tilde{N}}\phi_\perp,\\
&\delta_N\phi_a = \delta_{\mathbf{N}}\phi_a - \delta_{\tilde{N}}\phi_a = \delta_{\mathbf{N}}\phi_a - L_{\tilde{N}}\phi_a.
\end{aligned} \tag{5.10}$$

The coefficients

$$\underset{\sim}{\pi}^\perp(x) \equiv \epsilon\underset{\sim}{\eta}(x)\lambda^{\perp\perp}(x), \quad \underset{\sim}{\pi}^a(x) \equiv \underset{\sim}{\eta}(x)\lambda^{a\perp}(x) \tag{5.11}$$

of the directional derivatives $\delta_{\mathbf{N}}\phi_\perp$ and $\delta_{\mathbf{N}}\phi_a$ in the hypersurface Lagrangian $\delta_{\mathbf{N}}S^\phi[e]$ are identified with

the field momenta conjugate to $\phi_\perp(x)$ and $\phi_a(x)$. For convenience, we often supress the indices of the Levi-Civita form, using the space densities notation for the field momenta and the projections $\lambda^{\perp,\parallel}$,

$$\pi^\perp(x) \equiv \epsilon \underset{\sim}{\lambda}^{\perp\perp}(x), \quad \pi^a(x) \equiv \underset{\sim}{\lambda}^{a\perp}(x), \quad \underset{\sim}{\lambda}^{\perp,\parallel} \equiv g^{1/2}\lambda^{\perp,\parallel}. \tag{5.12}$$

The momenta (5.11) or (5.12) are the space tensor fields defined along the embedding, which are left unchanged by the Lie derivative $\mathbf{L}_{\tilde{N}}$ (Sec. I.7), and may thus be considered as hypertensors $\pi^\perp, \pi^\parallel$ canonically conjugate to the field hypertensors $\phi_\perp, \phi_\parallel$.

The hypersurface Lagrangian contains the terms depending on the derivatives of the lapse function [the last term in the expression (5.8)] and on the derivatives of the shift vector [arising when the Lie derivatives $L_{\tilde{N}}\phi_\perp$ and $L_{\tilde{N}}\phi_a$ are substituted from Eqs. (5.10) into the expression (5.8)]. In the fourth step, we integrate these terms by parts in the hypersurface Lagrangian (5.1), getting

$$\delta_{\mathbf{N}} S^\Phi = \int_m (\pi^\perp \delta_{\mathbf{N}} \phi + \pi^a \delta_{\mathbf{N}} \phi_a - N\mathring{H}^\Phi - N^a \mathring{H}^\Phi{}_a) + \int_{\partial m} (N\mathring{P}^{\Phi\, b} + N^a \mathring{P}^{\Phi\, b}{}_a)\, d\sigma_b. \tag{5.13}$$

This is the final form of the hypersurface Lagrangian. In it, we have introduced a number of abbreviations. Discussing the boundary term first,

$$\mathring{P}^{\Phi\, b} \equiv g^{-1/2}(\phi^b \pi^\perp - \epsilon\phi_\perp \pi^b), \quad \mathring{P}^{\Phi\, b}{}_a \equiv -g^{-1/2}\phi_a \pi^b \tag{5.14}$$

are space tensors which are bilinear forms of the field coordinates and the field momenta, and $d\sigma_b$ is a space-covector valued volume measure on the boundary. If the boundary is specified by the equations $x^a = \epsilon^a(\xi^A)$, where the intrinsic coordinates ξ^A, $A = 1, 2$, on ∂m are oriented so that the tangent vectors $\epsilon^a_A \equiv \epsilon^a{}_{,A}$ form with the outward normal ν^a to the boundary the right-handed system $\{\nu^a, e^a_A\}$, then

$$d\sigma_a = \eta_{abc}\epsilon^b_B \epsilon^c_C\, d\xi^B \wedge d\xi^C. \tag{5.15}$$

For a compact m, $\partial m = 0$ and the boundary term vanishes. In the rest of this paper, we will limit our attention to this case. The role of the boundary terms in an asymptotically flat spacetime is discussed, e.g., in Refs. 10.

The main part of the hypersurface Lagrangian is the space integral in Eq. (5.13). In it, we have introduced the super-Hamiltonian $\mathring{H}^\Phi$ and the supermomentum $\mathring{H}^\Phi$ of the field ϕ propagating on a given Riemannian background. In our notation, the symbols with a circle $^\circ$ above them denote the quantities taken on a fixed background, while the same symbols without the circle denote the corresponding quantities when the interaction of the field with geometry is included. Anticipating the results of the last paper of this series, this notation is used in the following formulas.

The super-Hamiltonian $\mathring{H}^\Phi$ and the supermomentum $\mathring{H}^\Phi{}_a$ are written in several parts, the interpretation of which is discussed in Secs. 6 and 8:

$$\mathring{H}^\Phi = \mathring{H}^\Phi_\dagger + \mathring{H}^\Phi_\ddagger, \tag{5.16}$$

$$\mathring{H}^\Phi_\dagger = H^\Phi_\dagger + 2K_{ab}P^{ab}, \tag{5.17}$$

$$H^\Phi_\dagger = \underset{\sim}{\Lambda}(\phi_\perp, \phi_a, \pi^\perp, \pi^a; \lambda^{\perp a}, \lambda^{ab}) - \epsilon\underset{\sim}{\lambda}^{\perp a}\phi_{\perp|a} - \underset{\sim}{\lambda}^{ab}\phi_{a|b}, \tag{5.18}$$

$$P^{ab} = \tfrac{1}{2}(-\phi^{(a}\underset{\sim}{\lambda}^{\perp b)} + \phi_\perp \underset{\sim}{\lambda}^{(ab)} - \phi^{(a}\pi^{b)}), \tag{5.19}$$

$$\mathring{H}^\Phi_\ddagger = H^\Phi_\ddagger = (\phi^a\pi^\perp - \epsilon\phi_\perp\pi^a)_{|a}, \tag{5.20}$$

$$\mathring{H}^\Phi{}_a = \phi_{\perp,a}\pi^\perp - 2\phi_{[a,b]}\pi^b - \phi_a \pi^b{}_{|b}. \tag{5.21}$$

The expression (5.19) is the space density form of the expression (5.9). The term $\mathring{H}^\Phi_\ddagger$ has arisen by the integration by parts of the terms containing the derivatives of the lapse function, and the term $\mathring{H}^\Phi{}_a$ has arisen similarly from the Lie derivatives $L_{\tilde{N}}\phi_\perp$ and $L_{\tilde{N}}\phi_a$.

The hypersurface Lagrangian (5.13)–(5.21) governs the dynamics of a covector field $\underline{\phi}$ on a fixed Riemannian background $(\mathcal{M}, \underline{g})$ expressed as the dynamics of conjugate hypervector fields $\{\phi_\perp, \phi_\parallel, \pi^\perp, \pi^\parallel\}$ in hyperspace. The way in which this is done is discussed in the next section.

6. HAMILTONIAN DYNAMICS OF SPACETIME HYPERTENSORS

In Sec. 4, we have found that the field equations may be obtained by varying the action functional

$$S^\Phi[\phi_{\perp,\parallel}, \lambda^{\perp,\parallel}; e] = \int dt\, \delta_{\mathbf{N}} S^\Phi \tag{6.1}$$

with respect to the projections $\phi_{\perp,\parallel}(Y)$ and $\lambda^{\perp,\parallel}(Y)$. The variation of the action functional (6.1) with respect to the foliation $e(Y)$ also yields a valid equation, which is a consequence of the field equations.

In Sec. 5, we have expressed the hypersurface Lagrangian $\delta_{\mathbf{N}} S^\Phi$ of a covector field in terms of the field projections $\phi_\perp(x)$, $\pi^\perp(x) = \epsilon g^{1/2}(x)\lambda^{\perp\perp}(x)$, $\phi_a(x)$, $\pi^a(x) = g^{1/2}(x)\lambda^{a\perp}(x)$, $\lambda^{\perp a}(x)$, $\lambda^{ab}(x)$, and the geometrical variables $g_{ab}(x)$ $K_{ab}(x)$, $N(x)$, $N^a(x)$. On a fixed Riemannian background $(\mathcal{M}, \underline{g})$, the geometrical variables become functionals of the foliation $e(Y)$. The general structure (5.13) of the hypersurface Lagrangian is preserved when we come to higher rank tensor fields $\underline{\phi}(X)$ (the tensor field of rank 2 is discussed in detail in Sec. 9). The covector field may thus serve as a typical illustration of the general case.

In the action (6.1), the directional derivatives $\delta_{\mathbf{N}}$ are taken along the deformation vectors tangent to the curve $e(t)$ in $\mathcal{E}$. Therefore,

$$\frac{d}{dt}\,\phi_\perp(Y) = \frac{d}{dt}\,\phi_\perp(x)[e(t)] = \delta_{\mathbf{N}}\phi_\perp(x)[e], \tag{6.2}$$

a similar equation holding for the projection $\phi_a(Y)$. The action (6.1) for the covector field then assumes the form

$$S^\Phi[\phi_\perp, \pi^\perp, \phi_a, \pi^a; \lambda^{\perp a}, \lambda^{a\perp}; e] = \int dt \int_m (\pi^\perp\dot{\phi}_\perp + \pi^a\dot{\phi}_a - N\mathring{H}^\Phi - N^a\mathring{H}^\Phi{}_a). \tag{6.3}$$

All variables in the action (6.3) are considered as functions of $Y \equiv (t, x)$. The action itself is in the Hamiltonian form, with $\phi_{\perp x}$, $\pi^{\perp x}$ and ϕ_{ax}, π^{ax} being the pairs of conjugate canonical variables, and $N^x\mathring{H}^\Phi{}_x + N^{ax}\mathring{H}^\Phi{}_{ax}$ playing the role of the Hamiltonian. Of course, the momenta $\pi^\perp(Y)$, $\pi^a(Y)$ can be varied in the action (6.3) in place of the projections $\lambda^{\perp\perp}(Y)$, $\lambda^{a\perp}(Y)$ because

they are functionals of these projections and of the embedding. In addition to the canonical variables, the action (6.3) depends through the super-Hamiltonian $\overset{\circ}{H}{}^{\Phi}$ on the variables $\lambda^{\perp a}(Y)$, $\lambda^{ab}(Y)$ as the Lagrange multipliers.

Varying the action (6.3) with respect to the canonical variables, we get the field equations in the Hamiltonian form. So, the variation of $\phi_\perp$ and $\pi^\perp$ gives

$$\begin{aligned}\dot{\phi}_\perp(x) &= [\phi_\perp(x),\ \overset{\circ}{H}{}^{\Phi}{}_{x'}]N^{x'} + [\phi_\perp(x),\ \overset{\circ}{H}{}^{\Phi}{}_{ax'}]N^{ax'},\\ \dot{\pi}^\perp(x) &= [\pi^\perp(x),\ \overset{\circ}{H}{}^{\Phi}{}_{x'}]N^{x'} + [\pi^\perp(x),\ \overset{\circ}{H}{}^{\Phi}{}_{ax'}]N^{ax'}.\end{aligned} \tag{6.4}$$

A similar pair of equations is obtained by varying ϕ_a and π^a. Using the arbitrariness of the foliation $e(t)$ and taking into account Eq. (6.2) for $\phi_\perp$ and $\pi^\perp$, we can rewrite Eqs. (6.4) as variational equations in hyperspace,

$$\begin{aligned}\delta_N \phi_\perp(x)[e] &= [\phi_\perp(x),\ \overset{\circ}{H}{}^{\Phi}{}_{x'}]N^{x'},\\ \delta_N \pi^\perp(x)[e] &= [\pi^\perp(x),\ \overset{\circ}{H}{}^{\Phi}{}_{x'}]N^{x'},\end{aligned} \tag{6.5}$$

and

$$\begin{aligned}\delta_{\tilde{N}} \phi_\perp(x)[e] &= [\phi_\perp(x),\ \overset{\circ}{H}{}^{\Phi}{}_{ax'}]N^{ax'},\\ \delta_{\tilde{N}} \pi^\perp(x)[e] &= [\pi^\perp(x),\ \overset{\circ}{H}{}^{\Phi}{}_{ax'}]N^{ax'}.\end{aligned} \tag{6.6}$$

These are the evolution equations we have already used in the kinematical considerations of Secs. II.4, II.5 and II.7. (See Ref. 2.)

Deriving Eqs. (6.4), we made a simplifying assumption that $e(t)$ is a foliation. This imposed a limitation $N(t,x) \neq 0$ on the lapse function. At this stage, however, we can easily take the limit in which $N(x)$ goes to zero at some points or regions of m and still get correct equations. Equations (6.4)–(6.6) thus hold for an arbitrary path in $\mathcal{E}$, not only for a foliation.

In addition to Hamilton's equations, supplementary equations are needed to determine the Lagrange multipliers $\lambda^{\perp b}$, λ^{ab}. The inspection of the hypersurface Lagrangian (5.13)–(5.21) reveals that the only place in which the multipliers enter into the action is the translational part $\overset{\circ}{H}{}^{\Phi}_{\uparrow}$ of the super-Hamiltonian. The supermomentum $\overset{\circ}{H}{}^{\Phi}{}_a$, the tilt part of the super-Hamiltonian $\overset{\circ}{H}{}^{\Phi}_{\swarrow}$, and the boundary terms (5.14) do not depend on the multipliers. Moreover, the translational part $\overset{\circ}{H}{}^{\Phi}_{\uparrow}(x)$ of the super-Hamiltonian is an algebraic function of $\lambda^{\perp b}(x)$ and $\lambda^{ab}(x)$. The equations obtained by varying the multipliers thus have the simple form

$$\frac{\partial \overset{\circ}{H}{}^{\Phi}_{\uparrow}(x)}{\partial \lambda^{\perp b}(x)} = 0 = \frac{\partial \overset{\circ}{H}{}^{\Phi}_{\uparrow}(x)}{\partial \lambda^{ab}(x)}. \tag{6.7}$$

We will use Eqs. (6.7) for the elimination of multipliers in Sec. 10.

Writing the Hamilton's equations as a double set of equations (6.5), (6.6), we have separated the normal evolution of the field from the tangential evolution. The tangential evolution is in a sense trivial, as it amounts to a mere reshuffling of the data on a single hypersurface expressed in terms of two equivalent embeddings, e and $e \circ \varphi$. This enabled us to derive the supermomentum $\overset{\circ}{H}{}^{\Phi}{}_a$ which generates the tangential dynamics from purely kinematical considerations in Sec. II.5. Indeed, the actual form (II.5.10) of the supermomentum of a covector field derived there coincides with the supermomentum (5.21) which we have arrived at by the rearrangement of the action functional S^Φ.

The normal evolution of the field under a hypersurface tilt is also reducible to purely kinematical considerations. Again, the actual form (II.4.5) of the tilt super-Hamiltonian $H^\Phi_\swarrow$ of a covector field which was derived in Sec. II.4 in a kinematical way, coincides with the tilt super-Hamiltonian (5.20) which emerges from the rearrangement of the spacetime action functional S^Φ.

The supermomentum and the tilt super-Hamiltonian are universal for a given tensor field, i.e., they depend only on the rank of that field, not on the specific dynamical properties of the field, which are coded in the Lagrangian potential Λ. The particular dynamics appears only in the translational part (5.18) of the super-Hamiltonian. A closer inspection of the expression (5.18) shows, however, that one part of that super-Hamiltonian, namely,

$$-\underset{\sim}{\lambda}{}^{ab}\phi_{a|b} - \epsilon \underset{\sim}{\lambda}{}^{\perp a}\phi_{\perp|a}, \tag{6.8}$$

is also universal. Is there any way of understanding the necessity of the terms (6.8) from kinematical considerations?

There is, and we are able to complete the kinematical argument now, having information which we lacked in Sec. II.4: we know that the momenta $\pi^\perp$ and π^a together with the multipliers $\lambda^{\perp b}$ and λ^{ab} compose the single spacetime tensor $\lambda^{\alpha\beta}$,

$$\begin{aligned}\lambda^{\alpha\beta} = {}& \lambda^{ab} e^\alpha_a e^\beta_b + \lambda^{\perp b} n^\alpha e^\beta_b + g^{-1/2}\pi^a e^\alpha_a n^\beta\\ & + \epsilon g^{-1/2}\pi^\perp n^\alpha n^\beta.\end{aligned} \tag{6.9}$$

Consequently, they must mix as the projections of a single spacetime tensor do under hypersurface tilts.

This means, according to Eqs. (II.3.16) applied to the projections of $\lambda^{\alpha\beta}$, that

$$\begin{aligned}\delta_\swarrow \pi^\perp &= -\epsilon\pi^c N_{,c} - \epsilon\underset{\sim}{\lambda}{}^{\perp c}N_{,c},\\ \delta_\swarrow \pi^a &= \pi^\perp N_{,a} - \underset{\sim}{\lambda}{}^{ac}N_{,c}.\end{aligned} \tag{6.10}$$

On the other hand, the same change should be generated by the super-Hamiltonian $\overset{\circ}{H}{}^{\Phi}$ according to the normal evolution equations (6.5). We call $\overset{\circ}{H}{}^{(\pi)}_{\swarrow}$ that part of the super-Hamiltonian $\overset{\circ}{H}{}^{\Phi}$ which is necessary for the purpose, and write

$$\begin{aligned}\delta_\swarrow \pi^\perp(x) &= [\pi^\perp(x),\ \overset{\circ}{H}{}^{(\pi)}_{\swarrow\, x'}]N^{x'}\\ &= -\frac{\delta \overset{\circ}{H}{}^{(\pi)}_{\swarrow\, x'}}{\delta\phi_\perp(x)}N^{x'},\\ \delta_\swarrow \pi^a(x) &= [\pi^a(x),\ \overset{\circ}{H}{}^{(\pi)}_{\swarrow\, x'}]N^{x'}\\ &= -\frac{\delta \overset{\circ}{H}{}^{(\pi)}_{\swarrow\, x'}}{\delta\phi_a(x)}N^{x'}.\end{aligned} \tag{6.11}$$

Comparing Eqs. (6.10) and (6.11) for an arbitrary tilt $N(x')$, proceeding in the same way as we did in Sec. II.4, we conclude that $\overset{\circ}{H}{}^{(\pi)}_{\swarrow}(x)$ is equal to the expression

$$\begin{aligned}\overset{\circ}{H}{}^{(\pi)}_{\swarrow} = H^{(\pi)}_{\swarrow} = {}& (\pi^\perp\phi^a{}_{|a} - \epsilon\pi^a\phi_{\perp|a})\\ & (-\epsilon\underset{\sim}{\lambda}{}^{\perp b}\phi_{\perp|b} - \underset{\sim}{\lambda}{}^{ab}\phi_{a|b}).\end{aligned} \tag{6.12}$$

The second line in Eq. (6.12) contains exactly the terms (6.8), the presence of which in $H^\Phi_\uparrow$ we have tried to

understand. They are necessary to generate the correct behavior of the field momenta under hypersurface tilts. The terms in the first line in Eq. (6.12) are also duely present in the full super-Hamiltonian $\overset{\circ}{H}{}^{\Phi}$; we have chosen, however, to include them into the tilt super-Hamiltonian $\overset{\circ}{H}{}^{\Phi}_{\mathcal{T}} = H^{\Phi}_{\mathcal{T}}$. Strictly speaking, this was not required. The tilt super-Hamiltonian $H^{\Phi}_{\mathcal{T}}$ was built to generate the tilt change of $\phi_\perp$ and ϕ_a and for this purpose, the expression

$$\overset{\circ}{H}{}^{(\Phi)}_{\mathcal{T}} \equiv H^{(\Phi)}_{\mathcal{T}} = \pi^{\perp}{}_{|a}\phi^a - \epsilon\,\pi^a{}_{|a}\phi_\perp \tag{6.13}$$

would be sufficient. We have included the terms $\pi^\perp\phi^a{}_{|a} - \epsilon\,\pi^a\phi_{\perp|a}$ into $\overset{\circ}{H}{}^{\Phi}_{\mathcal{T}}$ only to complete it nicely to a divergence.

The tilt super-Hamiltonians $H^{\Phi}_{\mathcal{T}}$ or $H^{(\Phi)}_{\mathcal{T}}$ explain merely the behavior of the canonical coordinates $\phi_\perp, \phi_a$ under the hypersurface tilts; to obtain a tilt super-Hamiltonian $H^{(\Phi,\Pi)}_{\mathcal{T}}$ which explains the behavior of the canonical momenta $\pi^\perp, \pi^a$ as well as of the canonical coordinates $\phi_\perp, \phi_a$, we should put

$$H^{(\Phi,\Pi)}_{\mathcal{T}} = H^{(\Phi)}_{\mathcal{T}} + H^{(\Pi)}_{\mathcal{T}}, \tag{6.14}$$

with $H^{(\Phi)}_{\mathcal{T}}$ and $H^{(\Pi)}_{\mathcal{T}}$ given by Eqs. (6.12) and (6.13). This leads to a finer splitting

$$\overset{\circ}{H}{}^{\Phi} = H^{\Phi} + 2P^{ab}K_{ab}, \tag{6.15}$$

$$H^{\Phi} = H^{(\Phi,\Pi)}_{\mathfrak{t}} + H^{(\Phi,\Pi)}_{\mathcal{T}} \tag{6.16}$$

of the field super-Hamiltonian $\overset{\circ}{H}{}^{\Phi}$ than that which was given in Sec. 5. Under this splitting,

$$H^{(\Phi,\Pi)}_{\mathfrak{t}} = \underset{\sim}{\Lambda}, \tag{6.17}$$

and the universal and the particular parts of the super-Hamiltonian are thus completely separated.

The coarser and, in a sense, less consistent splitting of Sec. 5 has, however, the advantages of its own. The tilt super-Hamiltonian $H^{\Phi}_{\mathcal{T}}$ is a complete divergence and, as we shall see in Sec. 8, it is directly related to a projection of the spin tensor $S^{\alpha\beta\gamma}$. Moreover, it does not depend on the multipliers $\lambda^{\perp b}$, λ^{ab} and drops thus completely out of Eq. (6.7). In the following, we shall use either one of the decompositions (6.12)–(6.17) or (5.16)–(5.20) according to convenience.

The canonical structure of the hypersurface action may be described in a more geometrical language. We have seen that the collection $\{\phi_\perp, \phi_{\shortparallel}; \pi^\perp, \pi^{\shortparallel}\}$ of hypertensors may be considered as a canonical coordinate chart in the phase space $\mathcal{P}$. After the coordinates $x^a = x^a(x)$ are introduced in m, the canonical coordinates in $\mathcal{P}$ are represented by a collection $\{\phi_\perp(x^b), \phi_a(x^b); \pi^\perp(x^b), \pi^a(x^b)\}$ of functions from $\mathcal{F}(\mathbf{R}^3)$. Let $\Pi \in \mathcal{P}$ be a point of $\mathcal{P}$ and $\mathbf{D}$ the exterior differential in $\mathcal{P}$. The canonical coordinates in $\mathcal{P}$ are to be considered as functions from $\mathcal{F}(\mathcal{P})$. Apply the exterior differentiation $\mathbf{D}$ to them and define the Cartan one-form $\Theta \in T^*_\Pi(\mathcal{P})$ by

$$\Theta = \pi^{\perp x}\,\mathbf{D}\,\phi_{\perp x} + \pi^{ax}\,\mathbf{D}\,\phi_{ax}. \tag{6.18}$$

The exterior differential of Θ yields then the symplectic form $\mathbf{\Omega} \in T^2_\Pi(\mathcal{P})$ on the phase space $\mathcal{P}$,

$$\mathbf{\Omega} = \mathbf{D}\Theta = \mathbf{D}\,\pi^{\perp x} \wedge \mathbf{D}\,\phi_{\perp x} + \mathbf{D}\pi^{ax} \wedge \mathbf{D}\,\phi_{ax}. \tag{6.19}$$

Let $\mathbf{\Gamma} = d\Gamma(t)/dt$ be the tangent vector to a curve $\Pi = \Gamma(t)$ in $\mathcal{P}$. We can form the inner product $\lrcorner$ in $\mathcal{P}$ between the tangent vector $\mathbf{\Gamma}$ and the one-form Θ,

$$\mathbf{\Gamma} \lrcorner \Theta = \pi^{\perp x}\,\dot{\phi}_{\perp x} + \pi^{ax}\,\dot{\phi}_{ax}. \tag{6.20}$$

This inner product enters the hypersurface action (6.3).

We have seen, however, that the single time parameter $t \in \mathbf{R}$ is most naturally replaced by the whole hypersurface $h \in \mathcal{H}$ in the relativistic field theory. The hyperspace $\mathcal{H}$ is an infinitely dimensional manifold and h thus has the character of a "many-fingered time."[11] Following this idea, it is appropriate to replace the single curve $\Pi = \Gamma(t)$ in $\mathcal{P}$ by a "many-fingered-time curve" in $\mathcal{P}$. Represent h by an embedding e which is a member of the equivalence class $\{e\}$ defining h. The many-fingered-time curve Γ is then defined as the mapping

$$\Gamma : e \in \mathcal{E} \to \Pi = \Gamma[e] \in \mathcal{P}. \tag{6.21}$$

The mapping (6.21) induces the linear mapping Γ_* of the tangent space $T_e(\mathcal{E})$ into the tangent space $T_{\Gamma[e]}(\mathcal{P})$,

$$\Gamma_* : \mathbf{N} \in T_e(\mathcal{E}) \to \mathbf{\Gamma}_{\mathbf{N}} \equiv \Gamma_*(\mathbf{N}) \equiv \delta_{\mathbf{N}}\,\Gamma[e] \in T_{\Gamma[e]}(\mathcal{P}), \tag{6.22}$$

which assigns to each tangent vector $\mathbf{N}$ from $T_e(\mathcal{E})$ the tangent vector $\mathbf{\Gamma}_{\mathbf{N}}$ from $T_{\Gamma[e]}(\mathcal{P})$. The inner product $\lrcorner$ of the Cartan's form Θ with this tangent vector $\mathbf{\Gamma}_{\mathbf{N}}$,

$$\mathbf{\Gamma}_{\mathbf{N}} \lrcorner \Theta = \int_m (\pi^\perp(x)\,\delta_{\mathbf{N}}\,\phi_\perp(x) + \pi^a(x)\,\delta_{\mathbf{N}}\,\phi_a(x)), \tag{6.23}$$

characterizes geometrically the canonical structure of the hypersurface Lagrangian (5.13).

7. GENERALIZED HAMILTONIAN DYNAMICS OF SPACETIME HYPERTENSORS

The Hamilton's equations (6.5), (6.6) (together with those for the canonical variables ϕ_a, π^a) and the supplementary equations (6.7) for the multipliers $\lambda^{\perp b}$, λ^{ab} are in themselves sufficient to determine the evolution of the covector field. However, the variation of the curve $e(t)$ also leads to a correct equation. The embedding e enters into the action (6.3) through the geometrical variables

$$\begin{aligned} &g_{ab}(x)[e] = g_{\alpha\beta}(e(x))\,e^\alpha_a e^\beta_b, \\ &K_{ab}(x)[e] = n_\gamma \frac{\nabla e^\gamma_a}{\partial x^b} = n_\gamma(e^\gamma_{a,b} + \Gamma^\gamma{}_{\alpha\beta}\,(e(x))\,e^\alpha_a e^\beta_b), \\ &N(x)[e] = \epsilon\,n_\alpha(x)[e]\,\dot{e}^\alpha(x), \quad N^a(x)[e] = e^a_\alpha(x)[e]\,\dot{e}^\alpha(x). \end{aligned} \tag{7.1, 7.2}$$

For a given $g_{\alpha\beta}(X)$ [and thus $\Gamma^\gamma{}_{\alpha\beta}(X)$], the intrinsic geometry $g_{ab}(x)[e]$ and the extrinsic curvature $K_{ab}(x)[e]$ depend only on the embedding $e(x)$, not on the further course of the $e(t)$ curve. The deformation $\mathcal{E}$-vector

$$(\mathbf{N})^{\alpha x} = \dot{e}^\alpha(x) \tag{7.3}$$

thus enters the action (6.3) linearly, through the lapse function and the shift vector (7.2). The momentum $p_\alpha(x)$ conjugate to $e^\alpha(x)$ does not thus depend on the deformation $\mathcal{E}$-vector at all,

$$p_\alpha(x) = -\epsilon\,n_\alpha(x)\,\overset{\circ}{H}{}^{\Phi}(x) - e^a_\alpha(x)\,\overset{\circ}{H}{}^{\Phi}{}_a(x). \tag{7.4}$$

Introducing it into the action (6.3), we cast the action into the homogeneous form

$$S^{\Phi}[\phi_\perp, \pi^\perp, \phi_a, \pi^a; \lambda^{\perp b}, \lambda^{ab}; e] = \int dt \int_m (p_\alpha \dot{e}^\alpha + \pi^\perp \dot{\phi}_\perp + \pi^a \dot{\phi}_a). \tag{7.5}$$

The action (7.5) is still to be varied with respect to the old variables, so that $p_\alpha(x)$ is to be varied under the constraint (7.4). Rather than taking the constraint in this form, we project it perpendicular and parallel to the embedding e, writing

$$\mathring{H}(x) = 0 = \mathring{H}_a(x), \tag{7.6}$$

with

$$\begin{aligned} \mathring{H}(x) &= \epsilon p_\perp(x) + \mathring{H}^{\Phi}(x), \\ \mathring{H}_a(x) &= p_a(x) + \mathring{H}^{\Phi}{}_a(x). \end{aligned} \tag{7.7}$$

We then add the constraints (7.7) to the action (7.5) by means of the new Lagrange multipliers $N(x)$, $N^a(x)$,

$$\begin{aligned} &S^{\Phi}[\phi_\perp, \pi^\perp, \phi_a, \pi^a; \lambda^{\perp b}, \lambda^{ab}; e^\alpha, p_\alpha; N, N^a] \\ &\quad \equiv \int dt \int_m (p_\alpha \dot{e}^\alpha + \pi^\perp \dot{\phi}_\perp + \pi^a \dot{\phi}_a - N\mathring{H} - N^a \mathring{H}_a), \end{aligned} \tag{7.8}$$

and vary all the variables $\phi_\perp$, $\pi^\perp$, ϕ_a, π^a, $\lambda^{\perp b}$, λ^{ab}, e^α, p_α, N, N^a independently.

The variation of the action (7.8) with respect to the momentum p_α gives the equation

$$\dot{e}^\alpha = N n^\alpha + N^a e^\alpha_a \tag{7.9}$$

which shows that the Lagrange multipliers N and N^a are actually the lapse function and the shift vector. This justifies the use of the old symbols N, N^a for the multipliers. The variables N, N^a, however, play an entirely different role in the old action (6.3) and in the new action (7.8). In the action (6.3), they are to be considered as the functionals (7.2) of the curve $e(t)$. In the action (7.8), they are treated as independent variables.

The variation of the action (7.8) with respect to the Lagrange multipliers N, N^a leads back to the constraints (7.6) [or (7.4)], and the variation with respect to e^α leads to the statement that the constraints (7.6) are preserved in time. The variation with respect to $\phi_\perp$, $\pi^\perp$, ϕ_a, π^a, $\lambda^{\perp b}$, λ^{ab} then gives the old field equations (6.4) and (6.7).

The inclusion of the variables e^α and p_α among the canonical variables leads necessarily to the constraints (7.6). The transition from the action (6.3) to the action (7.8) is sometimes called the *parametrization* of the action. It was discussed by Dirac[3] and ADM,[4] though primarily only on the flat background, for the insight which it provides to the constraint structure in pure geometrodynamics. Dirac called the newly introduced variables e^α and p_α the (hyper) surface variables. For further discussion of the parametrization of field theories, see also Ref. 12.

The parametrization of field theories has its predecessor in the parametrization of particle dynamics (see, e.g., Lanczos[13]). There, the time and energy are introduced as conjugate canonical variables. The Hamiltonian form of parametrized particle dynamics is called *generalized Hamiltonian dynamics* (see Estabrook and Wahlquist[14] for the history of this term and further references on the subject). Adopting this expression to the Hamiltonian dynamics described by the action (7.8), we will speak about generalized Hamiltonian dynamics of spacetime hypertensors.

The hypersurface Lagrangian corresponding to the parametrized form (7.8) of the action is

$$\begin{aligned} \delta_{\mathbf{N}} S^{\Phi} = \int_m (&p_\alpha(x) N^\alpha(x) + \pi^\perp(x) \delta_{\mathbf{N}} \phi_\perp(x) + \pi^a(x) \delta_{\mathbf{N}} \phi_a(x) \\ &- N(x) \mathring{H}(x) - N^a(x) \mathring{H}_a(x)). \end{aligned} \tag{7.10}$$

We can consider the spacetime covector–space density $p_\alpha(x)[e] \leftrightarrow p_{\alpha\, bcd}(x)[e]$ defined along the embedding e as the component expression of an $/\!/$-covector $\mathbf{p}$, and write the first term on the right-hand side of Eq. (7.10) as the inner product $\langle \mathbf{p}, \mathbf{N} \rangle$ in hyperspace. The $/\!/$-covector $\mathbf{p}$ has the components $p_\perp(x)[e]$ and $p_a(x)[e]$ with respect to the normal hyperbasis $\{\delta_{\perp x}, \delta_{ax}\}$. These are related to the energy and momentum densities measured by a family of observers moving in the normal direction to the hypersurface by the formulas derived in the next section.

Similarly as in the last section, the canonical structure of the hypersurface Lagrangian (7.10) can be described in geometrical terms. The collection $\{e, \phi_\perp, \phi_{/\!/}; \mathbf{p}, \pi^\perp, \pi^{/\!/}\}$ formed by the embedding e and the hypertensors $\phi_\perp, \phi_{/\!/}, \mathbf{p}, \pi^\perp, \pi^{/\!/}$ may be considered as a canonical coordinate chart in the generalized phase space ${}^e\mathcal{P}$. After the coordinates $x^a = x^a(x)$ and $X^\alpha = X^\alpha(X)$ are introduced in m and $/\!/$, the canonical coordinates in ${}^e\mathcal{P}$ are represented by the collection $\{e^\alpha(x^b), \phi_\perp(x^b), \phi_a(x^b); p_\alpha(x^b), \pi^\perp(x^b), \pi^a(x^b)\}$ of functions from $\mathcal{F}(\mathbf{R}^3)$. The canonical coordinates in ${}^e\mathcal{P}$ are to be considered as functions from $\mathcal{F}({}^e\mathcal{P})$. Let ${}^e\Pi \in {}^e\mathcal{P}$ be a point in the generalized phase space ${}^e\mathcal{P}$ and ${}^e\mathbf{D}$ the exterior differential in ${}^e\mathcal{P}$. Define the Cartan one-form ${}^e\Theta \in T^*_{e_\Pi}({}^e\mathcal{P})$ by

$${}^e\Theta = p_{\alpha x}\, {}^e\mathbf{D} e^{\alpha x} + \pi^{\perp x}\, {}^e\mathbf{D}_{\perp x} + \pi^{ax}\, {}^e\mathbf{D}\phi_{ax} \tag{7.11}$$

and the symplectic form ${}^e\Omega \in T^2_{\tilde{}}({}^e\mathcal{P})$ by

$${}^e\Omega = {}^e\mathbf{D}\, {}^e\Theta = {}^e\mathbf{D}p_{\alpha x} \wedge {}^e\mathbf{D}e^{\alpha x} + {}^e\mathbf{D}\pi^{\perp x} \wedge {}^e\mathbf{D}\phi_{\perp x} + {}^e\mathbf{D}^{ax} \wedge {}^e\mathbf{D}\phi_{ax}. \tag{7.12}$$

The many-fingered-time curve (6.21) in $\mathcal{P}$ may be lifted to ${}^e\mathcal{P}$,

$${}^e\Gamma : e \in \mathcal{E} \to {}^e\Gamma[e] \equiv \{e, \mathbf{p}[e],\ \Gamma[e]\} \in {}^e\mathcal{P}, \tag{7.13}$$

and the induced mapping ${}^e\Gamma_*$ from $T_e(\mathcal{E})$ into $T_{e_{\Gamma[e]}}({}^e\mathcal{P})$ defined. The tangent vector ${}^e\mathbf{\Gamma}_{\mathbf{N}} = {}^e\Gamma_*(\mathbf{N})$ from $T_{e_{\Gamma[e]}}({}^e\mathcal{P})$ enters into the inner product $\rfloor$ with the Cartan form ${}^e\Theta$ to yield the canonical structure of the hypersurface Lagrangian (7.10),

$$\begin{aligned} {}^e\mathbf{\Gamma}_{\mathbf{N}} \rfloor\, {}^e\Theta = \int_m (&p_\alpha(x) N^\alpha(x) + \pi^\perp(x)\, \delta_{\mathbf{N}} \phi_\perp(x) \\ &+ \pi^a(x)\, \delta_{\mathbf{N}} \phi_a(x)). \end{aligned} \tag{7.14}$$

Not the whole generalized phase space ${}^e\mathcal{P}$ is spanned by the dynamical trajectories of the system, because the generalized Hamiltonian dynamics takes place only on the constraint hypersurface in ${}^e\mathcal{P}$, defined by Eqs. (7.6). For this aspect of the generalized Hamiltonian dynamics, see again Ref. 14.

8. HYPERSURFACE LAGRANGIAN AND THE ENERGY-MOMENTUM TENSOR

In Sec. 3, we have derived the symmetrical energy–momentum tensor $T^{\alpha\beta}$ from the first-order spacetime form S^Φ of the action by the Belinfante–Rosenfeld procedure. We will now show how to obtain the projections $T^{\perp\perp}$, $T^{\perp}{}_a$, and T^{ab} of the symmetrical energy–momentum tensor directly from the hypersurface Lagrangian.

For this purpose, we rewrite the hypersurface action (6.3) in a slightly different form. We introduce into the hypersurface Lagrangian the extrinsic curvature

$$K_{ab}(x)[e] = -\frac{1}{2N}\,\delta_N g_{ab}(x) = \frac{1}{2N}\,(-\delta_{\mathbf{N}} g_{ab} + L_{\tilde{N}} g_{ab})$$

$$= \frac{1}{2N}\,(-\delta_{\mathbf{N}} g_{ab} + 2N_{(a|b)}) \tag{8.1}$$

from Eq. (I.9.7) and integrate the shift terms in the expression $-2NK_{ab}P^{ab}$ by parts. This brings us to the hypersurface Lagrangian

$$\begin{aligned}\delta_{\mathbf{N}} S^\Phi &= \int_m (\pi^\perp \delta_{\mathbf{N}}\phi_\perp + \pi^a \delta_{\mathbf{N}}\phi_a + P^{ab}\delta_N g_{ab} - NH^\Phi - N^a \mathring{H}^\Phi{}_a)\\ &= \int_m [\pi^\perp \delta_{\mathbf{N}}\phi_\perp + \pi^a \delta_{\mathbf{N}}\phi_a + P^{ab}\delta_{\mathbf{N}} g_{ab}\\ &\quad - NH^\Phi - N^a(\mathring{H}^\Phi{}_a - 2P^b_{a|b})]\end{aligned} \tag{8.2}$$

and to the corresponding hypersurface action S^Φ considered as a functional of the variables $\phi_\perp$, $\pi^\perp$, ϕ_a, π^a, $\lambda^{\perp b}$, λ^{ab}, and N, N^a, g_{ab}.

The symmetrical energy-momentum tensor $T^{\alpha\beta}(X)$ is given by the variational derivative (3.1) of the spacetime action S^Φ with respect to the spacetime metric $g_{\alpha\beta}(X)$. When expressing the action in the hypersurface form S^Φ, the new variables $\phi_\perp$, $\pi^\perp$, ϕ_a, π^a, $\lambda^{\perp b}$, λ^{ab}, and N, N^a, g_{ab} are the functionals of the old variables and of $g_{\alpha\beta}(X)$. The variational derivatives of the hypersurface action S^Φ with respect to the field variables $\phi_\perp$, $\pi^\perp$, ϕ_a, π^a, $\lambda^{\perp b}$, λ^{ab} vanish, due to the field equations. Modulo the field equations, the energy–momentum tensor $T^{\alpha\beta}$ may thus be expressed as

$$\begin{aligned}\tfrac{1}{2}|{}^4g|^{1/2}\,T^{\alpha\beta}(X) &= \frac{\delta S^\Phi}{\delta g_{\alpha\beta}(X)}\\ &= \frac{\delta S^\Phi}{\delta N^Y}\frac{\delta N^Y}{\delta g_{\alpha\beta}(X)} + \frac{\delta S^\Phi}{\delta N^{aY}}\frac{\delta N^{aY}}{\delta g_{\alpha\beta}(X)} + \frac{\delta S^\Phi}{\delta g_{abY}}\frac{\delta g_{abY}}{\delta g_{\alpha\beta}(X)}.\end{aligned} \tag{8.3}$$

The variational derivatives of N, N^a, and g_{ab} on a fixed embedding e_t with respect to the metric $g_{\alpha\beta}(X)$ of the surrounding spacetime are obtained by varying Eqs. (7.1), (7.2). The variation of the first of Eqs. (7.1) gives directly

$$\frac{\delta g_{ab}(Y)}{\delta g_{\alpha\beta}(X)} = e_a^{(\alpha} e_b^{\beta)}\,\delta(e(Y), X). \tag{8.4}$$

To get the variational derivatives of N^a and N, we first determine the variational derivatives of e^a_α and n_α. Varying the equation

$$e^a_\alpha = g^{ab} g_{\alpha\beta} e^\beta{}_{,b} \tag{8.5}$$

we get

$$\frac{\delta e^a_\alpha(Y)}{\delta g_{\mu\nu}(X)} = \epsilon\, e^{a(\mu} n^{\nu)} n_\alpha\,\delta(e(Y), X), \tag{8.6}$$

and varying the definition equations of n_α,

$$g^{\alpha\beta} n_\alpha n_\beta = \epsilon, \quad e^\alpha_a n_\alpha = 0, \tag{8.7}$$

we get

$$\frac{\delta n_\alpha(Y)}{\delta g_{\mu\nu}(X)} = \tfrac{1}{2}\epsilon\, n^\mu n^\nu n_\alpha\,\delta(e(Y), X). \tag{8.8}$$

Returning then to Eqs. (7.2), we obtain

$$\frac{\delta N^a(Y)}{\delta g_{\alpha\beta}(X)} = N e^{a(\alpha} n^{\beta)}\,\delta(e(Y), X), \tag{8.9}$$

$$\frac{\delta N(Y)}{\delta g_{\alpha\beta}(X)} = \tfrac{1}{2}\epsilon N n^\alpha n^\beta\,\delta(e(Y), X). \tag{8.10}$$

Due to the δ functions in Eqs. (8.4), (8.9), and (8.10), the integration over $Y \in \mathcal{N}$ in Eq. (8.3) is trivial. At this stage, it is straightforward to project the resulting equation into the $\perp\perp$, $\perp\parallel$, and $\parallel\parallel$ directions. Because $|{}^4g| = Ng^{1/2}$, we get

$$\underset{\sim}{T}{}^{\perp\perp}(Y) = \epsilon\,\frac{\delta S^\Phi}{\delta N(Y)}, \tag{8.11}$$

$$\underset{\sim}{T}{}_{\perp a}(Y) = \frac{\delta S^\Phi}{\delta N^a(Y)}, \tag{8.12}$$

$$N\underset{\sim}{T}{}^{ab}(Y) = 2\,\frac{\delta S^\Phi}{\delta g_{ab}(Y)}. \tag{8.13}$$

The variational derivatives (8.11), (8.12) are easily read off from the hypersurface action (6.1) generated by the hypersurface Lagrangian (8.2). We get

$$\underset{\sim}{T}{}^{\perp\perp}(x) = -\epsilon H^\Phi, \tag{8.14}$$

and

$$\underset{\sim}{T}{}_{\perp a}(x) = -\mathring{H}^\Phi{}_a + 2P^b_{a|b}. \tag{8.15}$$

To get the remaining variational derivative (8.13), we first evaluate the term

$$\begin{aligned}&\frac{\delta \int_N P^{mn}(Y)\,\delta_N g_{mn}(Y)}{\delta g_{ab}(Y)}\\ &= \frac{\delta \int_N (P^{mn}(Y)(\dot{g}_{mn}(Y) - L_{\tilde{N}} g_{mn}(Y))}{\delta g_{ab}(Y)}\\ &= \frac{\partial P^{mn}(Y)}{\partial g_{ab}(Y)}(\dot{g}_{mn}(Y) - L_{\tilde{N}} g_{mn}(Y)) - \dot{P}^{ab}(Y) + L_{\tilde{N}} P^{ab}(Y)\\ &= \left(-2\,\frac{\partial P^{mn}}{\partial g_{ab}} K_{mn} N - \delta_N P^{ab}\right).\end{aligned} \tag{8.16}$$

In the hypersurface Lagrangian (8.2), the supermomentum $\mathring{H}^\Phi{}_a$ does not depend on the intrinsic metric $g_{ab}(x)$. Therefore,

$$\begin{aligned}N\underset{\sim}{T}{}^{ab}(x) &= -4\,\frac{\partial P^{mn}}{\partial g_{ab}} K_{mn} N - 2\,\delta_N P^{ab}\\ &\quad - 2\,\frac{\delta}{\delta g_{ab}}\int_m NH^\Phi.\end{aligned} \tag{8.17}$$

The virtue of Eqs. (8.14), (8.15), and (8.17) is that they generate the projections of the symmetrical energy-momentum tensor directly from the hypersurface Hamiltonian $N^x \mathring{H}^\Phi{}_x + N^{ax}\mathring{H}^\Phi{}_{ax}$. We have derived them for a covector field $\underline{\phi}(X)$, but the derivation is equally valid for an arbitrary tensor field. According to Eq. (8.14), the H^Φ part of the super-Hamiltonian may be physically identified with the energy density on

e measured by a family of observers moving perpendicular to e. The super-Hamiltonian $\overset{\circ}{H}{}^{\Phi}$ of the field propagating on a Riemannian background thus differs from this energy density by the term $+2K_{ab}P^{ab}$. Similarly, the supermomentum $\overset{\circ}{H}{}^{\Phi}{}_{a}$ of the field propagating on a Riemannian background differs from the minus momentum density measured by a family of observers moving perpendicular to e by the term $2P^{b}_{alb}$, according to Eq. (8.15). In short,

$$\begin{aligned}\overset{\circ}{H}{}^{\Phi} &= -\epsilon T^{\perp\perp} + 2P^{ab}K_{ab},\\ \overset{\circ}{H}{}^{\Phi}{}_{a} &= -T_{\perp a} + 2P^{b}_{alb}.\end{aligned} \tag{8.18}$$

The symmetrical tensor density P^{ab}, introduced for the covector field by Eq. (5.9), enters prominently into all the formulas (8.17), (8.18). One can check directly that P^{ab} is nothing else but the projection $\frac{1}{2}g^{1/2}P^{ab\perp}$ of the spacetime tensor $P^{\alpha\beta\gamma}$ defined by Eq. (3.9), which participates in the construction of the symmetrical energy–momentum tensor (3.10). Indeed, for a covector field

$$P^{\alpha\beta\gamma} = \lambda^{(\alpha\beta)}\phi^{\gamma} - \lambda^{\gamma(\alpha}\phi^{\beta)} - \lambda^{(\alpha\gamma}\phi^{\beta)} \tag{8.19}$$

and the $ab\perp$ projection of $P^{\alpha\beta\gamma}$ leads to the expression (5.9),

$$P^{ab} = \tfrac{1}{2}g^{1/2}P^{ab\perp}. \tag{8.20}$$

Again, Eq. (8.20) is generally valid for an arbitrary tensor field $\underset{\sim}{\phi}(X)$.

The tensor $P^{\mu\nu\sigma}$ got to the formula (3.10) for the symmetrical energy–momentum tensor through the variation (3.7) of the spacetime affine connection. When the affine connection does not enter into the field Lagrangian, we say that the field has nonderivative gravitational coupling. This happens for the n-form fields, $n=0,1,2,3$, for which $\nabla\underset{\sim}{\phi} = \underset{\sim}{d}\phi$. The term $P^{\mu\nu\sigma}{}_{;\sigma}$ then disappears from the formula (3.10) for $T^{\mu\nu}$. Simultaneously, the extrinsic curvature K_{ab}, which got into the hypersurface Lagrangian by the rearrangement of the covariant derivatives $\lambda^{\{\alpha\}\beta}\phi_{\{\alpha\};\beta}$, disappears from there as well. Therefore,

$$P^{ab} = \tfrac{1}{2}\underset{\sim}{P}{}^{ab\perp} = 0 \tag{8.21}$$

for the fields with nonderivative gravitational coupling. The equations (8.14), (8.15), (8.17), (8.18) for the projections of the symmetrical energy–momentum tensor then vastly simplify, giving in this case

$$\begin{aligned}T^{\perp\perp} &= -\epsilon\overset{\circ}{H}{}^{\Phi} = -\epsilon H^{\Phi},\\ T_{\perp a} &= -\overset{\circ}{H}{}^{\Phi}{}_{a} = -H^{\Phi}{}_{a},\end{aligned} \tag{8.22}$$

and

$$NT^{ab} = -2\frac{\delta(N^{x'}\overset{\circ}{H}{}^{\Phi}{}_{x'})}{\delta g_{ab}} = -2\frac{\delta(N^{x'}H^{\Phi}{}_{x'})}{\delta g_{ab}}. \tag{8.23}$$

This is one of the reasons why the theories with nonderivative gravitational coupling are so much simpler than the theories with derivative gravitational coupling when expressed in the hyperspace language.

Similarly as P^{ab} may be identified with a projection of the spacetime tensor $P^{\alpha\beta\gamma}$, the tilt part H^{Φ}_{τ} of the super-Hamiltonian is related to the $\perp\perp a$ projection of the spin tensor $S^{\mu\nu\sigma}$, introduced by Eqs. (3.20) and (3.22):

$$H^{\Phi}_{\tau} = -\epsilon\underset{\sim}{S}{}^{\perp\perp a}{}_{|a}. \tag{8.24}$$

Check this again for the covector field. From Eq. (3.22),

$$S^{\mu\nu\sigma} = -\lambda^{[\nu\sigma]}\phi^{\mu} - \lambda^{\mu[\nu}\phi^{\sigma]} + \lambda^{[\sigma\mu}\phi^{\nu]}. \tag{8.25}$$

The $\perp\perp a$ projection of $S^{\mu\nu\sigma}$ gives the expression

$$S^{\perp\perp a} = -\epsilon(\pi^{\perp}\phi^{a} - \phi_{\perp}\pi^{a}), \tag{8.26}$$

so that Eqs. (8.26) and (5.20) lead to Eq. (8.24).

The $\Theta_{\perp}{}^{\perp}$ projection of the canonical energy–momentum tensor, however, does not coincide with the translational part of either $\overset{\circ}{H}{}^{\Phi}$ or H^{Φ}. When we project Eq. (3.20), taking into account the antisymmetry of $S^{\mu\nu\sigma}$ and applying the projection formulas of Sec. II.2 to the covariant derivatives of $S^{\mu\nu\sigma}$, we get

$$T^{\perp\perp} = \Theta^{\perp\perp} + S^{\perp\perp c}{}_{|c} - \epsilon S^{(ab)\perp}K_{ab}. \tag{8.27}$$

We have already identified $\underset{\sim}{T}{}^{\perp\perp}$ with $-\epsilon H^{\Phi}$ and $\underset{\sim}{S}{}^{\perp\perp c}{}_{|c}$ with $-\epsilon H^{\Phi}_{\tau}$, so that Eq. (8.27) can be written as

$$\underset{\sim}{\Theta}{}^{\perp\perp} = -\epsilon(H^{\Phi}_{\tau} - \underset{\sim}{S}{}^{(ab)\perp}K_{ab}). \tag{8.28}$$

Projecting the expression (8.25), we get

$$-\underset{\sim}{S}{}^{(ab)\perp} = \underset{\sim}{\lambda}{}^{(ab)}\phi_{\perp} - \lambda^{\perp(a}\phi^{b)}. \tag{8.29}$$

The term $-S^{(ab)\perp}K_{ab}$ thus differs from the term $+2P^{ab}K_{ab}$, which would be needed to turn $\Theta^{\perp\perp}$ into $-\epsilon\overset{\circ}{H}{}^{\Phi}_{\tau}$, by $\phi^{(a}\pi^{b)}K_{ab}$,

$$\Theta^{\perp\perp} = -\epsilon(\overset{\circ}{H}{}^{\Phi}_{\tau} + \phi^{(a}\pi^{b)}K_{ab}). \tag{8.30}$$

Finally, comparing the expression (8.26) with the first equation (5.14), we see that the boundary term $\overset{\circ}{P}{}^{\Phi b}$ coincides with the $\perp\perp b$ projection of the spin tensor,

$$\overset{\circ}{P}{}^{\Phi b} = -\epsilon S^{\perp\perp b}. \tag{8.31}$$

We have thus identified the different pieces of the hypersurface Lagrangian with various projections of the symmetrical energy–momentum tensor and the spin tensor. Also, which was the main task of this section, we have generated the T^{ab} projection of the symmetrical energy–momentum tensor directly from the hypersurface Lagrangian.

9. HYPERSURFACE LAGRANGIAN FOR SECOND-RANK TENSOR FIELDS

The higher rank tensor fields $\underset{\sim}{\phi}(X) \in \mathcal{T}^{0}_{n}(\mathcal{M})$ are easily handled by the algorithm developed in Sec. 5 for the covector field. We will outline the procedure and state the results for a second-rank covariant tensor field $\phi_{\alpha\beta}(X)$.

We start from the spacetime action

$$S^{\Phi} = \int_{\mathcal{M}} |{}^{4}g|^{1/2}(\lambda^{\alpha\beta\gamma}\phi_{\alpha\beta;\gamma} - \Lambda), \tag{9.1}$$

with

$$\Lambda = \Lambda(\phi_{\alpha\beta}, \lambda^{\alpha\beta\gamma}, g_{\alpha\beta}). \tag{9.2}$$

In the first step, we project the tensors $\phi_{\alpha\beta}$, $\lambda^{\alpha\beta\gamma}$, $g_{\alpha\beta}$ and express the Lagrangian potential in terms of the projections. The spacetime scalar character of Λ leads to certain identities [analogous to the identities (5.4)

and (5.6) for a covector field] which we will not, however, write down explicitly.

In the second step, we project the bilinear form $\lambda^{\alpha\beta\gamma}\phi_{\alpha\beta;\gamma}$,

$$\begin{aligned}\lambda^{\alpha\beta\gamma}\phi_{\alpha\beta;\gamma} &= \lambda^{abc}\phi_{ab;c} + \epsilon\lambda^{ab\perp}\phi_{ab;\perp} + \epsilon\lambda^{a\perp c}\phi_{a\perp;c} \\ &\quad + \epsilon\lambda^{\perp bc}\phi_{\perp b;c} + \lambda^{a\perp\perp}\phi_{a\perp;\perp} + \lambda^{\perp b\perp}\phi_{\perp b;\perp} \\ &\quad + \lambda^{\perp\perp c}\phi_{\perp\perp;c} + \epsilon\lambda^{\perp\perp\perp}\phi_{\perp\perp;\perp}, \end{aligned} \tag{9.3}$$

and use the projection formulas (II.2.7), (II.2.8) for the covariant derivatives of a second-rank tensor.

In the third step, we replace the normal directional derivatives δ_N by $\delta_{\mathbf{N}} - L_{\tilde{N}}$, similarly as we have done in Eqs. (5.10). We then identify the hyperfield momenta as the coefficients of $\delta_{\mathbf{N}}\phi_{\perp,\parallel}$ in the action:

$$\begin{aligned}&\pi^{\perp\perp} = g^{1/2}\lambda^{\perp\perp\perp}, \\ &\pi^{a\perp} = \epsilon g^{1/2}\lambda^{a\perp\perp}, \quad \pi^{\perp b} = \epsilon g^{1/2}\lambda^{\perp b\perp}, \\ &\pi^{ab} = g^{1/2}\lambda^{ab\perp}. \end{aligned} \tag{9.4}$$

The operations cast the hypersurface Lagrangian into the form

$$\begin{aligned}\delta_{\mathbf{N}}S^{\Phi} &= \int_m [(\pi^{\perp\perp}\delta_{\mathbf{N}}\phi_{\perp\perp} + \pi^{a\perp}\delta_{\mathbf{N}}\phi_{a\perp} \\ &\quad + \pi^{\perp b}\delta_{\mathbf{N}}\phi_{\perp b} + \pi^{ab}\delta_{\mathbf{N}}\phi_{ab})] \\ &\quad - (\pi^{\perp\perp}L_{\tilde{N}}\phi_{\perp\perp} + \pi^{a\perp}L_{\tilde{N}}\phi_{a\perp} + \pi^{\perp b}L_{\tilde{N}}\phi_{\perp b} \\ &\quad + \pi^{ab}L_{\tilde{N}}\phi_{ab}) + Q^a N_{,a} - 2NP^{ab}K_{ab} - NH_{\dagger}^{\Phi})], \end{aligned} \tag{9.5}$$

with

$$\begin{aligned}Q^a &= \pi^{\perp\perp}(\phi_{\perp}{}^a + \phi^a{}_{\perp}) - \epsilon\pi^{\perp a}\phi_{\perp\perp} - \epsilon\pi^{a\perp}\phi_{\perp\perp} \\ &\quad + \pi^{\perp b}\phi^a{}_b + \pi^{b\perp}\phi_b{}^a - \epsilon\pi^{ba}\phi_{b\perp} - \pi^{ab}\phi_{\perp b}, \end{aligned} \tag{9.6}$$

$$\begin{aligned}P^{ab} &= \tfrac{1}{2}(\epsilon\lambda^{\perp\perp(a}\phi^{b)}{}_{\perp} + \epsilon\lambda^{\perp\perp(a}\phi_{\perp}{}^{b)} - \epsilon\lambda^{(a\perp b)}\phi_{\perp\perp} - \epsilon\lambda^{\perp(ab)}\phi_{\perp\perp} \\ &\quad + \lambda^{c\perp(a}\phi_c{}^{b)} + \lambda^{\perp c(a}\phi^{b)}{}_c - \lambda^{(acb)}\phi_{\perp c} - \lambda^{c(ab)}\phi_{c\perp} \\ &\quad + \pi^{\perp(a}\phi_{\perp}{}^{b)} + \pi^{(a\perp}\phi^{b)}{}_{\perp} + \pi^{(ac}\phi^{b)}{}_c + \pi^{c(a}\phi_c{}^{b)}), \end{aligned} \tag{9.7}$$

and

$$\begin{aligned}H_{\dagger}^{\Phi} &= -\lambda^{\perp\perp c}\phi_{\perp\perp|c} - \epsilon\lambda^{\perp bc}\phi_{\perp b|c} - \epsilon\lambda^{a\perp c}\phi_{a\perp|c} \\ &\quad - \lambda^{abc}\phi_{ab|c} + \Lambda. \end{aligned} \tag{9.8}$$

In the fourth step, we integrate by parts the term $Q^a N_{,a}$ and those terms in the hypersurface Lagrangian (9.5) which contain the derivatives of the shift vector in $L_{\tilde{N}}\ldots$. This casts the hypersurface Lagrangian (9.5) into the final form

$$\begin{aligned}\delta_{\mathbf{N}}S^{\Phi} &= \int_m (\pi^{\perp\perp}\delta_{\mathbf{N}}\phi + \pi^{a\perp}\delta_{\mathbf{N}}\phi_{a\perp} + \pi^{\perp b}\delta_{\mathbf{N}}\phi_{\perp b} + \pi^{ab}\delta_{\mathbf{N}}\phi_{ab} \\ &\quad - N\overset{\circ}{H}{}^{\Phi} - N^a\overset{\circ}{H}{}^{\Phi}{}_a) + \int_{\partial m}(N\overset{\circ}{P}{}^{\Phi b} + N^a\overset{\circ}{P}{}^{\Phi b}{}_a)\,d\sigma_b. \end{aligned} \tag{9.9}$$

Similarly as in Sec. 5, we have introduced the abbreviations

$$\overset{\circ}{H}{}^{\Phi} = H^{\Phi} + 2P^{ab}K_{ab}, \tag{9.10}$$

$$H^{\Phi} = H_{\dagger}^{\Phi} + H_{\not\dagger}^{\Phi}, \tag{9.11}$$

$$H_{\not\dagger}^{\Phi} \equiv \overset{\circ}{H}{}_{\not\dagger}^{\Phi} = Q^a{}_{|a}, \tag{9.12}$$

$$\begin{aligned}\overset{\circ}{H}{}^{\Phi}{}_a &= \pi^{\perp\perp}\phi_{\perp\perp,a} \\ &\quad + \pi^{b\perp}\phi_{b\perp,a} - (\pi^{b\perp}\phi_{a\perp})_{,b} \\ &\quad + \pi^{\perp b}\phi_{\perp b,a} - (\pi^{\perp b}\phi_{\perp a})_{,b} \\ &\quad + \pi^{bc}\phi_{bc,a} - (\pi^{bc}\phi_{ac})_{,b} - (\pi^{bc}\phi_{ba})_{,c}, \end{aligned} \tag{9.13}$$

where Q^a, P^{ab}, and $H_{\dagger}^{\Phi}$ were already given by Eqs. (9.6)–(9.8). The supermomentum (9.13) and the tilt super-Hamiltonian (9.6), (9.12) again coincide with the expression (II.5.11) and (II.4.6) determined for the tensor field from purely kinematical considerations. All the relations (8.14), (8.15), (8.17)–(8.19), (8.22), (8.28) between the projections $T^{\perp\perp}$, $T_{\perp a}$, $\Theta_{\perp}{}^{\perp}$, $P^{ab\perp}$, $S^{\perp\perp a}$, and $S^{(ab)\perp}$ on one hand, and $\overset{\circ}{H}{}^{\Phi}$, $\overset{\circ}{H}{}^{\Phi}{}_a$, $H_{\not\dagger}^{\Phi}$, $H_{\dagger}^{\Phi}$ on the other hand, remain valid for the tensor field.

The boundary terms in the tensor field hypersurface Lagrangian (9.9) are given by

$$\begin{aligned}&\overset{\circ}{P}{}^{\Phi b} = Q^b, \\ &\overset{\circ}{P}{}^{\Phi b}{}_a = -(\pi^{b\perp}\phi_{a\perp} + \pi^{\perp b}\phi_{\perp a} + \pi^{bc}\phi_{ac} - \pi^{cb}\phi_{ca}). \end{aligned} \tag{9.14}$$

The hypersurface Lagrangian (9.9) may be specialized to symmetrical tensors. The symmetry conditions

$$\phi_{\alpha\beta} = \phi_{\beta\alpha}, \quad \lambda^{\alpha\beta\gamma} = \lambda^{\beta\alpha\gamma} \tag{9.15}$$

imply several symmetries of the projections:

$$\begin{aligned}&\phi_{ab} = \phi_{ba}, \quad \phi_{a\perp} = \phi_{\perp a}, \\ &\pi^{ab} = \pi^{ba}, \quad \pi^{a\perp} = \pi^{\perp a}, \\ &\lambda^{abc} = \lambda^{bac}, \quad \lambda^{a\perp c} = \lambda^{\perp ac}. \end{aligned} \tag{9.16}$$

Usings Eqs. (9.16), we get

$$H_{\dagger}^{\Phi} = -\underset{\sim}{\lambda}{}^{\perp\perp c}\phi_{\perp\perp|c} - 2\epsilon\underset{\sim}{\lambda}{}^{a\perp c}\phi_{a\perp|c} - \underset{\sim}{\lambda}{}^{(ab)c}\phi_{(ab)|c} + \underset{\sim}{\Lambda}, \tag{9.17}$$

$$H_{\not\dagger}^{\Phi} = 2(\pi^{\perp\perp}\phi_{\perp}^a - \epsilon\pi^{a\perp}\phi_{\perp\perp} + \pi^{b\perp}\phi_b^a - \epsilon\pi^{ab}\phi_{b\perp})_{|a}, \tag{9.18}$$

$$\begin{aligned}P^{ab} &= \epsilon\underset{\sim}{\lambda}{}^{(a\perp\perp}\phi^{b)}{}_{\perp} + \epsilon\underset{\sim}{\lambda}{}^{\perp\perp(a}\phi^{b)}{}_{\perp} - \epsilon\underset{\sim}{\lambda}{}^{\perp(ab)}\phi_{\perp\perp} + \underset{\sim}{\lambda}{}^{c\perp(a}\phi_c{}^{b)} \\ &\quad - \tfrac{1}{2}\underset{\sim}{\lambda}{}^{(ac)b}\phi_{c\perp} - \tfrac{1}{2}\underset{\sim}{\lambda}{}^{(bc)a}\phi_{c\perp} + \pi^{(a\perp}\phi^{b)}{}_{\perp} + \pi^{(ac}\phi^{b)}{}_c, \end{aligned} \tag{9.19}$$

$$\begin{aligned}\overset{\circ}{H}{}^{\Phi}{}_a &= \pi^{\perp\perp}\phi_{\perp\perp,a} + 2\pi^{b\perp}\phi_{b\perp,a} - 2(\pi^{b\perp}\phi_{a\perp})_{,b} \\ &\quad + \pi^{bc}\phi_{bc,a} - 2(\pi^{bc}\phi_{ac})_{,b}. \end{aligned} \tag{9.20}$$

In Sec. 11, we specialize Eqs. (9.6)–(9.13) further to the two-form sources with nonderivative gravitational coupling.

10. ELIMINATION OF λ MULTIPLIERS

The projections $\lambda^{\{\perp,\parallel\}\parallel}$ of the spacetime tensor $\lambda^{\{\alpha\}\beta}$ enter into the hypersurface Lagrangian as Lagrange multipliers. They are present only in the translational part $\overset{\circ}{H}{}_{\dagger}^{\Phi}$ of the super-Hamiltonian, not in the supermomentum $H^{\Phi}{}_a$, the tilt super-Hamiltonian $H_{\not\dagger}^{\Phi}$, or in the boundary terms. These statements, which hold for an arbitrary tensor field, can be checked directly for the covector field (with the multipliers $\lambda^{\perp b}$, λ^{ab}) by inspecting the formulas in Sec. 5, and for the second-rank tensor field (with the multipliers $\lambda^{\perp\perp c}$, $\lambda^{\perp bc}$, $\lambda^{a\perp c}$, λ^{abc}) by inspecting the analogous formulas in Sec. 9. We will discuss the elimination of the $\lambda^{\{\perp,\parallel\}\parallel}$ multipliers for the covector field, the generalization of the procedure for the higher rank tensor fields being straightforward.

The variation of the hypersurface action S^{Φ} with respect to the multipliers leads to Eqs. (6.7),

$$\frac{\partial\overset{\circ}{H}{}_{\dagger}^{\Phi}}{\partial\lambda^{\perp a}} = 0 = \frac{\partial\overset{\circ}{H}{}_{\dagger}^{\Phi}}{\partial\lambda^{ab}}. \tag{10.1}$$

Of course, we are free to replace $\overset{\circ}{H}{}_{\dagger}^{\Phi}$ in Eq. (10.1) by

$\mathring{H}^{\Phi}$, or even by $\mathring{H}$ of the generalized Hamiltonian dynamics, because the additional terms do not contain the multipliers.

From the actual form, (5.17)–(5.19), of $\mathring{H}^{\Phi}_{\uparrow}$ we can calculate the derivatives in Eq. (10.1) and get

$$g^{-1/2}\frac{\partial \mathring{H}^{\Phi}_{\uparrow}}{\partial \lambda^{\perp a}} = \frac{\partial \Lambda}{\partial \lambda^{\perp a}} - \epsilon\,\phi_{\perp a} - K_{ab}\phi^{b} = 0,$$
$$g^{-1/2}\frac{\partial \mathring{H}^{\Phi}_{\uparrow}}{\partial \lambda^{ab}} = \frac{\partial \Lambda}{\partial \lambda^{ab}} - \phi_{a|b} + \phi_{\perp}K_{ab} = 0. \tag{10.2}$$

In general, we have assumed that the equation $\partial\Lambda/\partial\lambda^{\alpha\beta} = \phi_{\alpha;\beta}$ is invertible with respect to $\lambda^{\alpha\beta}$. Consequently, the Lagrange multipliers $\lambda^{\perp a}$ and λ^{ab} may be determined from Eqs. (10.2) as algebraic functions of the canonical variables $\phi_{\perp}$, $\pi^{\perp}$, ϕ_a, π^a, the space derivatives $\phi_{a|b}$ and $\phi_{\perp|b}$ of canonical coordinates, and the geometrical variables g_{ab}, K_{ab}. Note that the Lagrange multipliers do not depend on the lapse function N or the shift vector N^a.

The multipliers determined from Eqs. (10.2) may then be introduced into the field equations. The elimination of the multipliers, however, is much more conveniently carried at the level of the action principle. The general transformation theory of the action functional allows us to eliminate in the action functional a part of the variables in terms of the rest of the variables by using the Euler equations of the variables which are to be eliminated. The action functional considered as a functional of the remaining variables then leads to the correct Euler equations under the variation of these variables. (See, e.g., Lanczos[13] for the discussion of this standard procedure.) In our case, the variables to be eliminated are $\lambda^{\perp b}$, λ^{ab}, and the Euler equations obtained by varying $\lambda^{\perp b}$, λ^{ab} are just the algebraic equations (10.1). Solving these equations for $\lambda^{\perp b}$, λ^{ab} and substituting these solutions into the translational super-Hamiltonian $\mathring{H}^{\Phi}_{\uparrow}$ (which is the only part of the hypersurface Lagrangian which contains the multipliers), we get the modified super-Hamiltonian

$$^{*}\mathring{H}^{\Phi}_{\uparrow} = {}^{*}\mathring{H}^{\Phi}_{\uparrow}[\phi_{\perp}, \pi^{\perp}, \phi_a, \pi^a, g_{ab}, K_{ab}]$$

and the modified action

$$^{*}S^{\Phi}[\phi_{\perp}, \pi^{\perp}, \phi_a, \pi^a, e] = \int dt \int_m (\pi^{\perp}\dot{\phi}_{\perp} + \pi^a\dot{\phi}_a - N^{*}\mathring{H}^{\Phi} - N^a\mathring{H}^{\Phi}{}_a), \tag{10.3}$$

which is equivalent to the original action S^{Φ}. The action (10.3) is again in the Hamiltonian form and the normal Hamilton's equations in hyperspace follow from it as in Sec. 6,

$$\delta_N\phi_{\perp}(x) = [\phi_{\perp}(x), {}^{*}\mathring{H}^{\Phi}{}_{x'}]N^{x'},$$
$$\delta_N\pi^{\perp}(x) = [\pi^{\perp}(x), {}^{*}\mathring{H}^{\Phi}{}_{x'}]N^{x'}, \tag{10.4}$$

with a similar set for the canonical pair ϕ_a, π^a. The tangential set (6.6) of Hamilton's equations is left untouched, because the elimination of the multipliers does not effect the supermomentum $\mathring{H}^{\Phi}{}_a$. The difference between Eqs. (6.5) and (10.4) is that the multipliers are kept fixed as external parameters in the Poisson brackets in Eqs. (6.5), but they are implicitly varied in Eqs. (10.4), because the canonical variables enter the modified Hamiltonian ${}^{*}\mathring{H}^{\Phi}$ also through the Lagrange multipliers. The equivalence of the two sets of equations, (6.5) and (10.4), is, however, ensured by Eqs. (10.1).

One can easily pass to the generalized Hamiltonian form of the modified action ${}^{*}S^{\Phi}$ by defining the modified super-Hamiltonian

$${}^{*}\mathring{H} = \epsilon p_{\perp} + {}^{*}\mathring{H}^{\Phi}, \tag{10.5}$$

following otherwise the procedures of Sec. 7.

To have a concrete example of the elimination process in mind, take the covector field described by the Lagrangian potential

$$\Lambda = -\tfrac{1}{2}\lambda^{\alpha\beta}\lambda_{\alpha\beta}. \tag{10.6}$$

This Lagrangian potential leads to the wave equation

$$\Box\phi_{\alpha} \equiv g^{\beta\gamma}\phi_{\alpha;\beta\gamma} = 0 \tag{10.7}$$

for the covector field. The vector potential ϕ_{α} of Maxwell's electrodynamics, of course, is subject to a different wave equation, containing the DeRham's D'Alembertian. We will study it in detail in the next section. Here, we have chosen the Lagrangian potential (10.6) consciously, in spite of the traditional difficulties to keep the energy positive and the spin 1 of the field pure, to illustrate the peculiarities of the elimination process for a field with derivative gravitational coupling.

We express the Lagrangian potential (10.6) in terms of the projected variables,

$$\Lambda = -\tfrac{1}{2}\lambda^{ab}\lambda_{ab} - \tfrac{1}{2}\epsilon\lambda^{\perp b}\lambda_{\perp b} - \tfrac{1}{2}\epsilon g^{-1}\pi^a\pi_a - \tfrac{1}{2}g^{-1}(\pi^{\perp})^2, \tag{10.8}$$

and determine the Lagrange multipliers from Eqs. (10.2),

$$\lambda_{\perp a} = -\phi_{\perp|a} - \epsilon K_{ab}\phi^{b}, \quad \lambda_{ab} = -\phi_{a|b} + \phi_{\perp}K_{ab}. \tag{10.9}$$

Substituting them into Eqs. (10.6), (5.18), (5.19), we get

$$^{*}H^{\Phi}_{\uparrow} = -\tfrac{1}{2}\epsilon g^{-1/2}\pi_a\pi^a + \tfrac{1}{2}g^{1/2}\phi^{a|b}\phi_{a|b} - \tfrac{1}{2}g^{-1/2}(\pi^{\perp})^2 + \tfrac{1}{2}\epsilon g^{1/2}\phi^{\perp|a}\phi_{\perp|a} - \tfrac{1}{2}\epsilon g^{1/2}K^{ab}K^{c}_{a}\phi_{b}\phi_{c} - \tfrac{1}{2}g^{1/2}K_{ab}K^{ab}(\phi_{\perp})^2 \tag{10.10}$$

and

$$^{*}P^{ab} = \tfrac{1}{2}(g^{1/2}\phi^{\perp|(a}\phi^{b)} - g^{1/2}\phi^{(a|b)}\phi^{\perp} - \pi^{(a}\phi^{b)} + [(\phi_{\perp})^2K^{ab} + K^{(ac}\phi_{c}\phi^{b)}]. \tag{10.11}$$

The super-Hamiltonian ${}^{*}\mathring{H}^{\Phi}_{\uparrow}$ is then obtained from Eq. (5.17). It is a quadratic function of the extrinsic curvature K_{ab}.

The elimination of multipliers for a field which is derivatively coupled to geometry screws the gravitational part of the super-Hamiltonian, as it makes the "supermetric" dependent on the source-field variables. We shall tell this part of the story in the final paper of this series.

11. HYPERSPACE DYNAMICS OF FIELDS WITH NONDERIVATIVE GRAVITATIONAL COUPLING

The presence of the extrinsic curvature K_{ab} in the hypersurface Lagrangian S^{Φ} is a source of numerous

complications, especially when the fields are to be dynamically coupled to the gravitational field. We have already seen in Sec. 8 that the extrinsic curvature drops out of the hypersurface Lagrangian for theories with nonderivative gravitational coupling. We shall study now such theories systematically.

We say that the field has nonderivative gravitational coupling when the Lagrangian L does not depend on the derivatives of the metric tensor $\underset{\sim}{g}$. This happens when the field $\underline{\phi}$ is an n-form field, $n=0,1,2,3$, $\underline{\phi}\in\mathcal{T}_n(\mathcal{M})$, and the Lagrangian depends on its derivatives solely through the exterior differential $\underline{d\phi}$. In other words, the tensors $\underline{\phi}$ and $\overline{\chi}$ in the first-order form of the action S^Φ should be completely antisymmetrical, so that

$$\overline{\chi}\rfloor\nabla\underline{\phi}=\overline{\chi}\rfloor\underline{d\phi}. \tag{11.1}$$

We will cast the action of the n-form fields with nonderivative gravitational coupling into hypersurface form, starting with the scalar field $\phi(X)\in\mathcal{F}(\mathcal{M})$ and paying special attention to the one-form field $\underline{\phi}(x)\in\mathcal{T}_1(\mathcal{M})$, which case includes the Maxwell electrodynamics.

A. Scalar field

The action S^Φ of a scalar field $\phi\in\mathcal{F}(\mathcal{M})$ is

$$S^\Phi=\int|{}^4g|^{1/2}(\lambda^\alpha\phi_{,\alpha}-\Lambda(\phi,\lambda^\alpha,g_{\alpha\beta})). \tag{11.2}$$

We project the bilinear form

$$|{}^4g|^{1/2}\lambda^\alpha\phi_{,\alpha}=Ng^{1/2}(\epsilon\lambda^\perp\phi_{,\perp}+\lambda^a\phi_{,a}) \tag{11.3}$$

and use the equation

$$\delta_{\mathbf{N}}\phi=\delta_N\phi-L_{\vec{N}}\phi=\epsilon\phi_{,\perp}N-\phi_{,a}N^a. \tag{11.4}$$

Identifying the scalar field momentum π with the coefficient of $\delta_{\mathbf{N}}\phi$ in the action (11.2),

$$\pi\equiv g^{1/2}\lambda^\perp, \tag{11.5}$$

we come to the hypersurface Lagrangian

$$\delta_{\mathbf{N}}S^\Phi=\int_m(\pi\,\delta_{\mathbf{N}}\phi-N\mathring{H}^\Phi-N^a\mathring{H}^\Phi{}_a), \tag{11.6}$$

with

$$\mathring{H}^\Phi=H^\Phi=g^{1/2}(-\lambda^a\phi_{,a}+\Lambda(\phi,\pi,\lambda^a,g_{ab})), \tag{11.7}$$

$$\mathring{H}^\Phi{}_a=\pi\phi_{,a}. \tag{11.8}$$

We see that K_{ab} does not occur in the field super-Hamiltonian (11.7), and P^{ab} for the scalar field thus vanishes. This implies the simplified equations (8.22), (8.23) for the projections of the symmetrical energy–momentum tensor.

Let us write yet the condition that $\Lambda(\phi,\pi,\lambda^a,g_{ab})$ behaves as a spacetime scalar under hypersurface tilts. We have

$$\delta_\tau\pi=-g^{1/2}\lambda^aN_{,a},\quad \delta_\tau\lambda^a=\epsilon\pi g^{-1/2}N^{,a} \tag{11.9}$$

by Eqs. (II.3.1), while

$$\delta_\tau g_{ab}=0,\quad \delta_\tau\phi=0. \tag{11.10}$$

Consequently, the condition $\delta_\tau\Lambda=0$, expressing the scalar character of Λ, becomes

$$g^{1/2}\frac{\partial\Lambda}{\partial\pi}\lambda_a=\epsilon g^{-1/2}\frac{\partial\Lambda}{\partial\lambda^a}\pi. \tag{11.11}$$

This is an analog of a more complicated equation (5.6) for the covector field.

B. n-form fields, n = 1, 2, 3

To specialize the general results of Sec. 5 to a one-form field with nonderivative gravitational coupling, we impose the condition that $\lambda^{\alpha\beta}$ is an antisymmetrical tensor,

$$\lambda^{(\alpha\beta)}=0,\quad \lambda^{\alpha\beta}=\lambda^{[\alpha\beta]}. \tag{11.12}$$

Projected, Eq. (11.12) informs us that λ^{ab} is to be treated as an anitsymmetrical tensor,

$$\lambda^{(ab)}=0,\quad \lambda^{ab}=\lambda^{[ab]}, \tag{11.13}$$

and that

$$\lambda^{\perp a}=-\pi^a,\quad \pi^\perp=0. \tag{11.14}$$

As a consequence of Eqs. (11.13), (11.14), P^{ab} defined by Eq. (5.19) vanishes,

$$P^{ab}=0. \tag{11.15}$$

Further, the different parts of the super-Hamiltonian $\mathring{H}^\Phi$, defined by Eqs. (5.18), (5.20), (6.12), (6.13), (6.17) reduce to

$$\begin{aligned}&\mathring{H}^\Phi_\tau=H^\Phi_\tau=-\epsilon(\phi_\perp\pi^a)_{|a},\\ &\mathring{H}^\Phi_\perp=H^\Phi_\perp=-\underset{\sim}{\lambda}{}^{ab}\phi_{[a,b]}+\epsilon\pi^a\phi_{\perp,a}+\underset{\sim}{\Lambda},\end{aligned} \tag{11.16}$$

$$\begin{aligned}&\mathring{H}^{(\Phi)}_\tau=-\epsilon\phi_\perp\pi^a{}_{|a},\quad \mathring{H}^{(\tau)}_\tau=-\underset{\sim}{\lambda}{}^{ab}\phi_{[ab]},\\ &\mathring{H}^{(\Phi,\tau)}_\perp=\underset{\sim}{\Lambda}(\phi_a,\pi^a;\phi,\lambda^{ab};g_{ab}).\end{aligned} \tag{11.17}$$

Note that the space derivatives of $\phi_\perp$ dropped out from Eqs. (11.17) and consequently also from the total super-Hamiltonian

$$\mathring{H}^\Phi=H^\Phi=-\underset{\sim}{\lambda}{}^{ab}\phi_{[a,b]}-\epsilon\phi_\perp\pi^a{}_{|a}+\underset{\sim}{\Lambda}. \tag{11.18}$$

Similarly, the supermomentum (5.21) reduces to

$$\mathring{H}^\Phi{}_a=-\phi_a\pi^b{}_{|b}-2\phi_{[a,b]}\pi^b. \tag{11.19}$$

It does not depend on $\phi_\perp$ at all.

The hypersurface Lagrangian is to be considered as a functional of the variables ϕ_a, π^a, and $\phi_\perp$, $\lambda^{ab}=\lambda^{[ab]}$. We have seen that $\phi_\perp$ enters $\delta_{\mathbf{N}}S^\Phi$ in a purely algebraic way solely through the super-Hamiltonian (11.18), so that it stands on an equal footing with the λ^{ab} multiplier. The variables $\pi^\perp$, $\lambda^{\perp b}$ were eliminated from the action altogether, using Eqs. (11.14). The hypersurface action is thus in a canonical form in the pair of variables ϕ_a, π^a, and it depends on $\phi_\perp$ and $\lambda^{ab}=\lambda^{[ab]}$ as Lagrange multipliers,

$$S^\Phi=S^\Phi[\phi_a,\pi^a;\phi_\perp,\lambda^{ab}]=\int dt\int_m(\pi^a\phi_a-N\mathring{H}^\Phi-N^a\mathring{H}^\Phi{}_a). \tag{11.20}$$

The Euler equations following from the action (11.20) thus split into Hamilton's equations for the canonical variables ϕ_a, π^a, and the multiplier equations

$$g^{-1/2}\frac{\partial\mathring{H}^\Phi}{\partial\lambda^{ab}}=-\phi_{[a,b]}+\frac{\partial\Lambda}{\partial\lambda^{[ab]}}=0 \tag{11.21}$$

and

$$\frac{\partial\mathring{H}^\Phi}{\partial\phi_\perp}=-\epsilon\pi^a{}_{|a}+\frac{\partial\underset{\sim}{\Lambda}}{\partial\phi_\perp}=0. \tag{11.22}$$

In general, $3+1$ equations (11.21), (11.22) may be solved for the $3+1$ multipliers $\lambda^{ab}=\lambda^{[ab]}$ and $\phi_\perp$. In special cases, however, the equations (11.21) and (11.22) for λ^{ab}, $\phi_\perp$ are not all independent and lead then to the constraints for the dynamical variables ϕ_a, π^a. This happens in an important case of Maxwell's electrodynamics, which we will discuss in a little while.

The two-form fields with nonderivative gravitational coupling may be treated by specializing the results of Sec. 9 to completely antisymmetrical tensors $\phi_{\alpha\beta}$, $\lambda^{\alpha\beta\gamma}$,

$$\phi_{(\alpha\beta)}=0, \quad \lambda^{(\alpha\beta)\gamma}=\lambda^{\alpha(\beta\gamma)}=0. \tag{11.23}$$

Projecting Eqs. (11.23), we get the conditions

$$\phi_{\perp\perp}=0, \quad \phi_{\perp a}=-\phi_{a\perp}, \quad \phi_{(ab)}=0, \tag{11.24}$$

$$\pi^{\perp\perp}=0, \quad \pi^{\perp a}=\pi^{a\perp}=0, \quad \pi^{(ab)}=0, \tag{11.25}$$

$$\lambda^{\perp\perp c}=0, \quad \lambda^{\perp ab}=-\lambda^{a\perp b}=g^{-1/2}\pi^{ab}, \quad \lambda^{(ab)c}=\lambda^{a(bc)}=0. \tag{11.26}$$

Due to Eqs. (11.24)–(11.26), the hypersurface action may be expressed as a functional of the antisymmetrical tensors ϕ_{ab}, π^{ab}, λ^{abc}, and the covector $\phi_{\perp a}$. The hypersurface action has the canonical form with respect to the variables ϕ_{ab}, π^{ab}, while $\phi_{\perp a}$ and λ^{abc} play the role of Lagrange multipliers. The conditions (11.24)–(11.26) imply that $P^{ab}=0$ and the super-Hamiltonian $\overset{\circ}{H}{}^{\Phi}$ and the supermomentum $\overset{\circ}{H}{}^{\Phi}{}_a$ reduce to the form

$$\overset{\circ}{H}{}^{\Phi}=H^{\Phi}=-g^{1/2}\lambda^{abc}\phi_{ab,c}-2\epsilon\,\phi_{\perp b}\,\pi^{ab}{}_{,a} + g^{1/2}\Lambda(\phi_{ab},\pi^{ab};\phi_{\perp b},\lambda^{abc}), \tag{11.27}$$

$$\overset{\circ}{H}{}^{\Phi}{}_a=H^{\Phi}{}_a=\pi^{bc}\phi_{bc,a}-2(\pi^{bc}\phi_{ac})_{,b}. \tag{11.28}$$

Note that $\phi_{\perp b}$ enters the super-Hamiltonian (11.27) algebraically and it is not to be found in the supermomentum. The conclusions to be drawn from Eqs. (11.27), (11.28) are parallel to those for the one-forms.

Following the regularities which appear in the construction of $\overset{\circ}{H}{}^{\Phi}$ and $\overset{\circ}{H}{}^{\Phi}{}_a$ for the one-form fields [Eqs. (11.18), (11.19)] and two-form fields [Eqs. (11.27), (11.28)], we easily guess what the $\overset{\circ}{H}{}^{\Phi}$ and $\overset{\circ}{H}{}^{\Phi}{}_a$ are for the three-form fields.

For three-forms, the antisymmetrical tensors ϕ_{abc}, π^{abc} play the role of canonical variables, and the antisymmetrical tensor $\phi_{\perp ab}$, plays the role of a Lagrange multiplier. The super-Hamiltonian and supermomentum are given by the expressions

$$\overset{\circ}{H}{}^{\Phi}=H^{\Phi}=-3\epsilon\,\phi_{\perp ab}\,\pi^{abc}{}_{,c}+g^{1/2}\Lambda, \tag{11.29}$$

$$\overset{\circ}{H}{}^{\Phi}{}_a=\pi^{bcd}\phi_{bcd,a}-3(\pi^{bcd}\phi_{acd})_{,b}. \tag{11.30}$$

The first-order formalism for the four-forms becomes completely degenerate, because the antisymmetrical tensor $\lambda^{\alpha\beta\gamma\delta\epsilon}$ of the rank 5 automatically vanishes. The n-forms game thus stops with $n=3$.

C. Proca's field and Maxwell's electrodynamics

We illustrate the general theory of one-form fields with nonderivative gravitional coupling by choosing the special Lagrangian

$$L=-\phi_{[\alpha,\beta]}\phi^{[\alpha,\beta]}-\tfrac{1}{2}\mu^2\phi_\alpha\phi^\alpha. \tag{11.31}$$

For $\mu\neq 0$, the Lagrangian (11.31) describes neutral vector bosons with mass (Compton wavelength) μ; for $\mu=0$, we pass to Maxwell's electrodynamics. The field equations

$$2\phi^{[\alpha,\beta]}{}_{;\beta}-\mu^2\phi^\alpha=0 \tag{11.32}$$

imply (for $\mu\neq 0$) the Lorentz condition

$$\phi^\alpha{}_{;\alpha}=0 \tag{11.33}$$

which ensures the positive definite character of the field energy and its pure spin 1 character. The differential operator acting on ϕ^α in the first term of Eq. (11.34) is the DeRham's d'Alembertian.

Defining $\lambda^{\alpha\beta}$ in the standard way,

$$\lambda^{\alpha\beta}=\frac{\partial L}{\partial\phi_{\alpha;\beta}}=\frac{\partial L}{\partial\phi_{[\alpha,\beta]}}=-2\phi^{[\alpha,\beta]}=\lambda^{[\alpha\beta]}. \tag{11.34}$$

We see it is necessarily antisymmetrical, because L depends only on the antisymmetrical combination $\phi_{[\alpha;\beta]}=\phi_{[\alpha,\beta]}$ of the covariant derivatives $\phi_{\alpha;\beta}$. This is a situation which we have already mentioned in Sec. 2. Introducing $\lambda^{\alpha\beta}$ into the action by means of the Legendre dual transformation, we arrive at the first-order action with the Lagrangian potential (2.8),

$$\Lambda=-\tfrac{1}{4}\lambda^{\alpha\beta}\lambda_{\alpha\beta}+\tfrac{1}{2}\mu^2\phi_\alpha\phi^\alpha. \tag{11.35}$$

The Lagrangian potential (11.35) differs from the Lagrangian potential (10.6) by the mass term and by the explicit assumption that $\lambda^{\alpha\beta}$ in Eq. (11.35) is an antisymmetrical tensor. Expressed in terms of the hypertensors ϕ_a, π^a, $\phi_\perp$, λ^{ab}, the Lagrangian potential (11.35) becomes

$$\overset{\circ}{H}{}_1^{(\phi,\pi)}=g^{1/2}\Lambda=-\tfrac{1}{4}g^{1/2}\lambda_{ab}\lambda^{ab}-\tfrac{1}{2}\epsilon g^{-1/2}\pi_a\pi^a + \tfrac{1}{2}\mu^2 g^{1/2}(\phi_a\phi^a+\epsilon\,\phi_\perp{}^2), \tag{11.36}$$

the full super-Hamiltonian being given by Eq. (11.18).

The Lagrange multipliers λ^{ab}, $\phi_\perp$ may be determined from Eqs. (11.21), (11.22),

$$\lambda_{ab}=-2\phi_{[a,b]}\equiv B_{ab}, \tag{11.37}$$

$$\phi_\perp=-\mu^{-2}g^{-1/2}\pi^a{}_{|a}. \tag{11.38}$$

B_{ab} is the analog of the magnetic field strength. Eliminating the multipliers λ^{ab}, $\phi_\perp$ from the action, we get the modified action ${}^*S^\Phi$, with

$${}^*\overset{\circ}{H}{}^{\Phi}=-\tfrac{1}{2}\epsilon g^{-1/2}\pi_a\pi^a+\tfrac{1}{4}g^{1/2}B_{ab}B^{ab}+\tfrac{1}{2}\mu^2 g^{1/2}\phi_a\phi^a - \tfrac{1}{2}\epsilon\mu^{-2}g^{-1/2}(\pi^a{}_{|a})^2, \tag{11.39}$$

and

$${}^*\overset{\circ}{H}{}^{\Phi}{}_a=-\phi_a\pi^b{}_{|b}+B_{ab}\pi^b. \tag{11.40}$$

An alternative approach would be to eliminate only the λ^{ab} multipliers from the action S^Φ, but leave $\phi_\perp$ as a multiplier in the super-Hamiltonian.

For $\mu=0$, we pass to Maxwell's electrodynamics. The canonical momentum π^a is the electric field strength measured by an observer moving perpendicular to the hypersurface. The λ equations (11.21) still lead to the identification (11.37) of λ_{ab} with the magnetic

field strength B_{ab}. The $\phi_\perp$ equation (11.22), however, leads to the divergence constraint

$$\pi^a{}_{|a} = 0 \tag{11.41}$$

for the electric field strength. We thus lack the equation which would enable us to determine the scalar potential $\phi_\perp$. We must leave it in the super-Hamiltonian, eliminating only the λ^{ab} multipliers,

$$*\mathring{H}^\Phi = *H^\Phi = -\epsilon\,\phi_\perp \pi^a{}_{|a} - \tfrac{1}{2}\epsilon g^{-1/2}\pi_a\pi^a + \tfrac{1}{4} g^{1/2} B_{ab} B^{ab}. \tag{11.42}$$

The term $-\epsilon\,\phi_\perp \pi^a{}_{|a}$ is often set apart, the remainder being the standard expression for the electromagnetic field energy in terms of the electric and magnetic field strengths. The two expressions, with and without the term $-\epsilon\,\phi_\perp \pi^a{}_{|a}$, are numerically equal to each other, modulo the constraint (11.41). Similarly, one often drops the term $-\phi_a \pi^b{}_{|b}$ from the supermomentum (11.40), identifying the rest with the Poynting vector. These modifications of the action, achieved by using the Euler equation (11.41) back in the action, are entirely permissible. However, the truncated $*\mathring{H}^\Phi$ and $*\mathring{H}^\Phi{}_a$, with the terms containing $\pi^a{}_{|a}$ omitted, do not close according to the universal relations (12.45)–(12.47) which we derive in the next section. The right-hand sides of Eqs. (12.45)–(12.47) written for the truncated $*\mathring{H}^\Phi$, $*\mathring{H}^\Phi{}_a$ contain some extra terms proportional to $\pi^a{}_{|a}$. For principal reasons, it is thus better to work with the original expressions (11.42) and (11.40).

12. CLOSING OF CONSTRAINTS IN GENERALIZED HAMILTONIAN DYNAMICS

The super-Hamiltonian $\mathring{H}$ and the supermomentum $\mathring{H}_a$ of the generalized Hamiltonian dynamics are constrained to vanish. In order that the constraints (7.6) be preserved along a path in $\mathcal{E}$, the Poisson brackets of the constraint functions must close. In a first-order theory, the presence of the additional equations (6.7) for the λ multipliers complicates the closing relations. After the multipliers are eliminated, the Poisson brackets between the modified constraint functions $*\mathring{H}$ and $*\mathring{H}_a$ close in the universal way which is independent of the tensor character of the field. In this section, we derive the closing relations of the constraint functions $\mathring{H}$, $\mathring{H}_a$ and, by eliminating the λ multipliers, pass to the closing relations of the modified constraint functions $*\mathring{H}$, $*\mathring{H}_a$.

We discuss the closing relations first for the scalar field $\phi(X) \in \mathcal{F}(\mathcal{M})$. The generalization to tensor fields is then straightforward.

Our starting point is the fact discussed already in Sec. 4, that the hypersurface action S^Φ does not depend on the foliation $e^\alpha(Y)$ if the spacetime fields $\phi(X)$, $\lambda^\alpha(X)$ are kept fixed. For our purposes, it is best to write the hypersurface action in the form

$$S^\Phi[\phi, \pi; \lambda^a; e^\alpha] = \int dt \int_m (\pi\dot{\phi} - N^\alpha \mathring{H}^\Phi{}_\alpha), \tag{12.1}$$

with

$$N^\alpha = \dot{e}^\alpha, \qquad \mathring{H}^\Phi{}_\alpha = \epsilon n_\alpha \mathring{H}^\Phi + e^a_\alpha \mathring{H}^\Phi{}_a. \tag{12.2}$$

Its foliation independence leads to the identity

$$\delta_e S^\Phi = \frac{\delta S^\Phi}{\delta \phi_Y}\,\delta_e \phi_Y + \frac{\delta S^\Phi}{\delta \pi^Y}\,\delta_e \pi^Y + \frac{\delta S^\Phi}{\delta \lambda^{aY}}\,\delta_e \lambda^{aY} + \frac{\delta S^\Phi}{\delta e^{\alpha Y}}\,\delta e^{\alpha Y} \equiv 0, \tag{12.3}$$

where δ_e are the variations of the hyperfields induced by the variation $\delta e^\alpha(Y)$ of the foliation when the spacetime fields $\phi(X)$, $\lambda^\alpha(X)$, and $g_{\alpha\beta}(X)$ are kept fixed [compare with the variations (4.15)].

The identity (12.3) remains an identity if the time derivatives $\dot{\phi}$ and $\dot{\pi}$ are eliminated from it by means of the Hamilton's equations

$$\frac{\delta S^\Phi}{\delta \phi_Y} = 0 = \frac{\delta S^\Phi}{\delta \pi^Y}. \tag{12.4}$$

In this case, it reduces to the statement

$$\left[\frac{\delta S^\Phi}{\delta \lambda^{aY}}\,\delta_e \lambda^{aY} + \frac{\delta S^\Phi}{\delta e^{\alpha Y}}\,\delta e^{\alpha Y}\right]_{\dot{\phi},\dot{\pi}\ \text{eliminated}} \equiv 0. \tag{12.5}$$

Because $\mathring{H}^\Phi{}_\alpha$ depends algebraically on λ^a, and the derivatives $\partial \mathring{H}^\Phi{}_\alpha / \partial \lambda^a$ do not depend on the velocities $\dot{\phi}$, $\dot{\pi}$, the first term in Eq. (12.5) becomes

$$\left[\frac{\delta S^\Phi}{\delta \lambda^{aY}}\,\delta_e \lambda^{aY}\right]_{\dot{\phi},\dot{\pi}\ \text{eliminated}} = \frac{\delta S^\Phi}{\delta \lambda^{aY}}\,\delta_e \lambda^{aY} = -\int dt \int_{x \in m} \frac{\partial \mathring{H}^\Phi{}_\alpha}{\partial \lambda^a}(x)\,\delta_e \lambda^a(x)\, N^\alpha(x). \tag{12.6}$$

The rearrangement of the second term in Eq. (12.5) is more complicated. First,

$$\frac{\delta S^\Phi}{\delta e^{\alpha Y}}\,\delta e^{\alpha Y} = -\frac{\delta \mathring{H}^\Phi{}_{\alpha Y}}{\delta e^{\beta Y'}} N^{\alpha Y} \delta e^{\beta Y'} - \mathring{H}_{\alpha Y'}\,\delta \dot{e}^{\alpha Y'} = -\frac{\delta \mathring{H}^\Phi{}_{\alpha Y}}{\delta e^{\beta Y'}} N^{\alpha Y} \delta e^{\beta Y'} + (\mathring{H}^\Phi{}_{\beta Y'})^{\cdot}\,\delta e^{\beta Y'}. \tag{12.7}$$

Second, we evaluate the last term in Eq. (12.7) using the Hamilton equations,

$$(\mathring{H}^\Phi{}_{\beta Y'})^{\cdot}\,\delta e^{\beta Y'} = \int dt\,(\mathring{H}^\Phi{}_{\beta x'}(t))^{\cdot}\,\delta e^{\beta x'}(t) = \int dt \left\{(\mathring{H}^\Phi{}_{\beta x'}, \mathring{H}^\Phi{}_{\alpha x}) N^{\alpha x}\,\delta e^{\beta x'} + \frac{\delta \mathring{H}^\Phi{}_{\beta x'}}{\delta e^{\alpha x}} N^{\alpha x}\,\delta e^{\beta x'} + \left(\frac{\partial \mathring{H}^\Phi{}_\beta}{\partial \lambda^a}\dot{\lambda}^a\right)_x \delta e^{\beta x}\right\}. \tag{12.8}$$

Third, we use the fact that $\mathring{H}^\Phi{}_\alpha$ does not depend on the time derivatives of e^β, so that

$$\frac{\delta \mathring{H}_{\alpha Y'}}{\delta e^{\beta Y}} = \frac{\delta \mathring{H}^\Phi{}_\alpha(x')}{\delta e^\beta(x)}\,\delta(t', t). \tag{12.9}$$

Last, we put all Eqs. (12.7)–(12.9) together, noticing that neither $\mathring{H}^\Phi{}_\alpha$ nor its (variational) derivatives depend on $\dot{\phi}$, $\dot{\pi}$,

$$\left[\frac{\delta S^{\Phi}}{\delta e^{\alpha Y}}\,\delta e^{\alpha Y}\right]_{\dot{\phi},\dot{\pi}\ \text{eliminated}} = \int dt\left\{\left(\frac{\delta \mathring{H}^{\Phi}{}_{\beta x'}}{\delta e^{\alpha x}} - \frac{\delta \mathring{H}^{\Phi}{}_{\alpha x}}{\delta e^{\beta x'}}\right) + [\mathring{H}^{\Phi}{}_{\beta x'}, \mathring{H}^{\Phi}{}_{\alpha x}]\right\} N^{\alpha x}\delta e^{\beta x'} + \int dt\left(\frac{\partial \mathring{H}^{\Phi}{}_{\beta}}{\partial \lambda^{a}}\dot{\lambda}^{a}\right)_{x'}\delta e^{\beta x'}. \tag{12.10}$$

The expression in the $\{\}$ bracket is equal to the Poisson bracket between the constraint functions

$$\mathring{H}_{\alpha} = p_{\alpha} + \mathring{H}^{\Phi}{}_{\alpha} \tag{12.11}$$

of the generalized Hamiltonian dynamics,

$$[\mathring{H}_{\beta x'}, \mathring{H}_{\alpha x}] = \frac{\delta \mathring{H}^{\Phi}{}_{\beta x'}}{\delta e^{\alpha x}} - \frac{\delta \mathring{H}^{\Phi}{}_{\alpha x}}{\delta e^{\beta x'}} + [\mathring{H}^{\Phi}{}_{\beta x'}, \mathring{H}^{\Phi}{}_{\alpha x}]. \tag{12.12}$$

The constraint functions $\mathring{H}_{\alpha}$ may also replace the functions $\mathring{H}^{\Phi}{}_{\alpha}$ in the derivatives $\partial \mathring{H}^{\Phi}{}_{\alpha}/\partial\lambda^{a}$. Substituting Eqs. (12.6), (12.10), (12.12) into the identity (12.5), we get

$$-\int dt\,[\mathring{H}_{\alpha x}, \mathring{H}_{\beta x'}]N^{\alpha x}\,\delta e^{\beta x'} + \int_{m} dt\,\frac{\partial \mathring{H}_{\alpha}}{\partial \lambda^{a}}\left(\dot{\lambda}^{a}\,\delta e^{\alpha} - N^{\alpha}\,\delta_{e}\lambda^{a}\right) \equiv 0. \tag{12.13}$$

For symmetry, we write

$$\mathbf{M} = \delta_{e} \tag{12.14}$$

for the $/\!/$-vector characterizing the change of the foliation $e(Y)$, remembering that $\mathbf{N}$ is the tangent $/\!/$-vector to the foliation itself. In this notation,

$$(\dot{\lambda}^{a}\,\delta e^{\alpha} - N^{\alpha}\delta_{e}\lambda^{a}) = (M^{\alpha}\delta_{\mathbf{N}} - N^{\alpha}\delta_{\mathbf{M}})\,\lambda^{a},$$

and the identity (12.13) assumes the final form

$$\int dt\left\{[\mathring{H}_{\alpha x}, \mathring{H}_{\beta x'}] + \left(\frac{\partial \mathring{H}_{\alpha}}{\partial \lambda^{a}}(x)\,\frac{\delta\lambda^{a}(x)}{\delta e^{\beta}(x')} - (\alpha x \longleftrightarrow \beta x')\right)\right\} N^{\alpha x}M^{\alpha x'} \equiv 0. \tag{12.15}$$

Equation (12.15) expresses the fact that the action S^{Φ} does not depend on the foliation $e(Y)$ when the spacetime fields $\phi(X)$, $\lambda^{\alpha}(X)$, and $g_{\alpha\beta}(X)$ are kept fixed.

Due to the arbitrariness of the $/\!/$-vectors $\mathbf{N}$ and $\mathbf{M}$, Eq. (12.15) implies that

$$\mathring{H}_{\alpha x\,\beta x'} + \mathring{I}_{\alpha x\,\beta x'} \equiv 0, \tag{12.16}$$

where we have introduced the abbreviations

$$\mathring{H}_{\alpha x\,\beta x'} \equiv [\mathring{H}_{\alpha x}, \mathring{H}_{\beta x'}], \qquad \mathring{I}_{\alpha x\,\beta x'} \equiv \frac{\partial \mathring{H}_{\alpha}}{\partial \lambda^{a}}(x)\,\frac{\delta\lambda^{a}(x)}{\delta e^{\beta}(x')} - (\alpha x \longleftrightarrow \beta x'). \tag{12.17}$$

Our next task is to project Eq. (12.16) into $\perp x\ \perp x'$, $ax\ \perp x'$, and $ax\ bx'$ directions. The projections of $\mathring{H}_{\alpha x\,\beta x'}$ are the same for all tensor fields, being equal to

$$\mathring{H}_{\perp x\,\perp x'} = [\mathring{H}(x), \mathring{H}(x')] + \epsilon(\mathring{H}^{a}(x)\,\delta_{,a}(x,x') - (x \longleftrightarrow x')), \tag{12.18}$$

$$\mathring{H}_{ax\,\perp x'} = \epsilon[\mathring{H}_{a}(x), \mathring{H}(x')] - \epsilon\mathring{H}(x)\,\delta_{,a}(x,x'), \tag{12.19}$$

$$\mathring{H}_{ax\,bx'} = [\mathring{H}_{a}(x), \mathring{H}_{b}(x')] - (\mathring{H}_{b}(x)\,\delta_{,a}(x,x') - (ax \longleftrightarrow bx')). \tag{12.20}$$

We shall prove Eq. (12.18) and indicate how to proceed in the case of Eqs. (12.19)–(12.20). The basic properties of the Poisson brackets give

$$\mathring{H}_{\perp x\,\perp x'} = [\mathring{H}(x), \mathring{H}(x')] - [n^{\alpha}(x), \mathring{H}(x')]\mathring{H}_{\alpha}(x) - [\mathring{H}(x), n^{\alpha}(x')]\mathring{H}_{\alpha}(x'). \tag{12.21}$$

Because $\mathring{H}(x')$ does not depend on p_{α}, and $n^{\alpha}(x)$ does not depend on the field variables ϕ, π,

$$[n^{\alpha}(x), \mathring{H}(x')] = [n^{\alpha}(x), \epsilon p_{\perp}(x')] = \epsilon\,\delta_{\perp x'}\,n^{\alpha}(x). \tag{12.22}$$

The derivative $\delta_{\perp x'}\,n^{\alpha}(x)$ is given by Eq. (I.6.16),

$$\delta_{\perp x'}\,n^{\alpha}(x) = -\,e^{a\alpha}(x)\,\delta_{,a}(x,x') + A^{\alpha}_{\perp}(x)\,\delta(x,x'). \tag{12.23}$$

Substituting Eq. (12.23) into Eq. (12.22), and Eq. (12.22) into Eq. (12.21), the term $A^{\alpha}_{\perp}(x)\,\delta(x,x')$ drops out and we get the desired result, Eq. (12.18).

The proof of Eqs. (12.19) and (12.20) is similar. It uses Eqs. (I.6.9), (I.6.10), and the relation $L_{\tilde{N}}n^{\alpha} = \delta_{\tilde{N}}n^{\alpha}$ for the hypertensor n^{α}.

Returning to the projections of $\mathring{I}_{\alpha x\,\beta x'}$, we get

$$\mathring{I}_{\perp x\,\perp x'} = \epsilon\,\frac{\partial \mathring{H}}{\partial \lambda^{a}}(x)\,\delta_{\perp x'}\,\lambda^{a}(x) - (x \longleftrightarrow x'), \tag{12.24}$$

$$\mathring{I}_{ax\,\perp x'} = -\,\epsilon\,\frac{\partial \mathring{H}}{\partial \lambda^{b}}(x')\,\delta_{bx}\lambda^{a}(x'), \tag{12.25}$$

and

$$\mathring{I}_{ax\,bx'} = 0, \tag{12.26}$$

because only $\mathring{H}$ depends on λ^{a}.

In Eq. (12.24), we can substitute for $\delta_{\perp x'}\lambda^{a}(x)$ the tilt change $\delta_{\not\perp x'}\lambda^{a}(x)$, because the translational change drops out due to the interchange $(x \longleftrightarrow x')$. Because

$$\delta_{\not\perp x'}\lambda^{a}(x) = \lambda^{\perp}(x)\,\delta_{,}{}^{a}(x,x') = g^{-1/2}(x)\,\pi(x)\,\delta_{,}{}^{a}(x,x'), \tag{12.27}$$

Eq. (12.24) becomes

$$\mathring{I}_{\perp x\,\perp x'} = \epsilon\,\mathring{I}^{a}(x)\,\delta_{,a}(x,x') - (x \longleftrightarrow x'), \qquad \mathring{I}^{a} \equiv g^{-1/2}\,\pi\,g^{ab}\,\frac{\partial \mathring{H}}{\partial \lambda^{b}}. \tag{12.28}$$

In Eq. (12.25), $\delta_{bx}\lambda^{a}(x')$ is determined from the relation

$$\delta_{\tilde{N}}\lambda^{a}(x) = L_{\tilde{N}}\lambda^{a}(x) = \lambda^{a}{}_{,b}N^{b} - \lambda^{b}N^{a}{}_{,b} \tag{12.29}$$

for the hypertensor λ^{a}; it is thus given by

$$\delta_{bx}\lambda^{a}(x') = \lambda^{a}{}_{,b}(x')\,\delta(x',x) - \delta^{a}_{b}\lambda^{c}(x')\,\delta_{,c'}(x',x). \tag{12.30}$$

Collecting Eqs. (12.18)–(12.20), and Eqs. (12.25), (12.26), (12.28), (12.30) together, the projected identities (12.16) assume the final form

$$[\mathring{H}(x), \mathring{H}(x')] = -\,\epsilon\{(\mathring{H}^{a}(x) + \mathring{I}^{a}(x))\,\delta_{,a}(x,x') - (x \longleftrightarrow x')\}, \tag{12.31}$$

$$[\mathring{H}_a(x),\ \mathring{H}(x')]=\mathring{H}(x)\,\delta_{,a}(x,x')+\frac{\partial\mathring{H}}{\partial\lambda^a}(x')\,\lambda^a{}_{,b}(x')\,\delta(x',x)$$
$$-\frac{\partial\mathring{H}}{\partial\lambda^b}(x')\,\lambda^c(x')\,\delta_{,c}(x',x), \tag{12.32}$$

$$[\mathring{H}_a(x),\ \mathring{H}_b(x')]=\mathring{H}_b(x)\,\delta_{,a}(x,x')-(ax\longleftrightarrow bx'). \tag{12.33}$$

The closing relations (12.31)–(12.33) ensure that the constraints

$$\mathring{H}(x)=0=\mathring{H}_a(x) \tag{12.34}$$

will hold on a deformed embedding, if they hold on the original embedding together with the λ equations

$$\frac{\partial\mathring{H}}{\partial\lambda^a}(x)=0. \tag{12.35}$$

Indeed,

$$\delta_{\mathbf{N}}\mathring{H}(x)=[\mathring{H}(x),\ \mathring{H}_{x'}]N^{x'}+[\mathring{H}(x),\ \mathring{H}_{ax'}]N^{ax'}$$
$$+\frac{\partial\mathring{H}}{\partial\lambda^a}(x)\,\delta_{\mathbf{N}}\lambda^a(x), \tag{12.36}$$

and Eqs. (12.34), (12.35) imply through the closing relations (12.31), (12.32) that $\delta_{\mathbf{N}}\mathring{H}(x)=0$. The same argument applies to $\delta_{\mathbf{N}}\mathring{H}_a(x)$.

The λ equations (12.35) resemble the constraints (12.34) in the respect that they contain no directional derivatives $\delta_{\mathbf{N}}$ and therefore limit only the choice of the field variables λ^a, ϕ, π on the initial hypersurface. Superficially, one tends to guess that the Poisson brackets

$$\left[\frac{\partial\mathring{H}}{\partial\lambda^a}(x),\ \mathring{H}(x')\right]\quad\text{and}\quad\left[\frac{\partial\mathring{H}}{\partial\lambda^a}(x),\ \mathring{H}_b(x')\right] \tag{12.37}$$

must also be closed, in order that the λ equations be preserved along a curve in $\mathcal{E}$. Such a conjecture, however, is misleading. On the contrary, we will prove that the Poisson brackets (12.37) *cannot* be closed; if they were, the theory would be internally inconsistent.

To see this, calculate

$$\delta_{\mathbf{N}}\left(\frac{\partial\mathring{H}}{\partial\lambda^a}(x)\right)=\frac{\partial^2\mathring{H}}{\partial\lambda^a\partial\lambda^b}(x)\,\delta_{\mathbf{N}}\lambda^b(x)$$
$$+\left[\frac{\partial\mathring{H}}{\partial\lambda^a}(x),\ \mathring{H}_{x'}\right]N^{x'}+\left[\frac{\partial\mathring{H}}{\partial\lambda^a}(x),\ \mathring{H}_{ax'}\right]N^{ax'} \tag{12.38}$$

The left-hand side of this equation vanishes, because the λ equations hold on an arbitrary embedding. Equation (12.38) may then be used for determining $\lambda^a(x)$ on the deformed embedding if $\lambda^a(x)$ is known on the old embedding; it is a linear nonhomogeneous equation for $\delta_{\mathbf{N}}\lambda^a(x)$ which has the unique solution due to the assumed regularity of the $\partial^2\mathring{H}/\partial\lambda^a\partial\lambda^b$ matrix. However, if the Poisson brackets (12.37) were closed, the constraints (12.34) and the λ equations (12.35) on the initial embedding would imply that this solution is necessarily zero, $\delta_{\mathbf{N}}\lambda^a(x)=0$, which is an obvious contradiction, because $\lambda^a(x)$ may change from one embedding to another. Therefore, the Poisson brackets between $\partial\mathring{H}/\partial\lambda^a$ and the constraint functions $\mathring{H}$, $\mathring{H}_a$ cannot close.

This is a vital difference between the constraints (12.34) and the λ equations (12.35).

Generalize now the closing relations (12.31)–(12.33) to a covector field. For the covector field, $\mathring{I}_{\alpha x\,\beta x'}$ in the identity (12.16) assumes the form

$$\mathring{I}_{\alpha x\,\beta x'}=\frac{\partial\mathring{H}_\alpha}{\partial\lambda^{\perp b}}(x)\,\frac{\delta\lambda^{\perp b}(x)}{\delta e^\beta(x')}+\frac{\partial\mathring{H}_\alpha}{\partial\lambda^{ab}}(x)\,\frac{\delta\lambda^{ab}(x)}{\delta e^\beta(x')}-(x\longleftrightarrow x'). \tag{12.39}$$

The $\perp x\ \perp x'$ projection of the expression (12.39) gives

$$\mathring{I}_{\perp x\,\perp x'} \tag{12.40}$$
$$=\epsilon\,\frac{\partial\mathring{H}}{\partial\lambda^{\perp b}}(x)\,\delta_{\perp x'}\lambda^{\perp b}(x)+\epsilon\,\frac{\partial\mathring{H}}{\partial\lambda^{ab}}(x)\,\delta_{\perp x'}\lambda^{ab}(x)-(x\longleftrightarrow x').$$

Again, due to the antisymmetrization $(x\longleftrightarrow x')$, we can replace the expressions $\delta_{\perp x'}\lambda^{\perp b}(x)$ and $\delta_{\perp x'}\lambda^{ab}(x)$ in Eq. (12.40) by the tilt changes, which we read off from Eqs. (II.3.16),

$$\delta_{\perp x'}\lambda^{\perp b}(x)=-\epsilon\lambda^{cb}(x)\,\delta_{,c}(x,x')+\lambda^{\perp\perp}(x)\,\delta_{,}{}^{b}(x,x')$$
$$=-\epsilon\lambda^{cb}(x)\,\delta_{,c}(x,x')+\epsilon g^{-1/2}\pi^{\perp}(x)\,\delta_{,}{}^{b}(x,x'),$$

$$\delta_{\perp x'}\lambda^{ab}(x)=\lambda^{a\perp}\delta_{,}{}^{b}(x,x')+\lambda^{\perp b}\,\delta_{,}{}^{a}(x,x')$$
$$=\pi^a\,\delta_{,}{}^{b}(x,x')+\lambda^{\perp b}\,\delta_{,}{}^{a}(x,x'). \tag{12.41}$$

This leads to Eq. (12.28) and the closing relation (12.31), with $\mathring{I}^a(x)$ given by

$$\mathring{I}_a=\epsilon g^{-1/2}\pi^{\perp}\frac{\partial\mathring{H}}{\partial\lambda^{\perp a}}-\epsilon\lambda_a{}^{b}\,\frac{\partial\mathring{H}}{\partial\lambda^{\perp b}}$$
$$+g^{-1/2}\pi^b\frac{\partial\mathring{H}}{\partial\lambda^{ba}}+\frac{\partial\mathring{H}}{\partial\lambda^{ab}}\,\lambda^{\perp b}. \tag{12.42}$$

Similarly, the $ax\ \perp x'$ projection of the expression (12.40) leads to the closing relation which is analogous to Eq. (12.32), with appropriate Lie derivative terms. The closing relation (12.33) remains unchanged.

At the end, let us pass to the closing relations among the constraint functions

$$*\mathring{H}(x)=\mathring{H}(x)[\phi,\pi,\lambda^a(\phi,\pi,e^\alpha),\ e^\alpha,p_\alpha],$$
$$*\mathring{H}_a(x)=\mathring{H}_a(x)[\phi,\pi,\ e^\alpha,p_\alpha] \tag{12.43}$$

of the modified Hamiltonian theory in which the λ multipliers were eliminated by means of the λ equations (12.35). When we calculate the Poisson brackets between the constraint functions (12.43), any derivative of the original functions $\mathring{H}(x)$, with respect to the variables ϕ, π, e^α hidden in the multipliers $\lambda^a[\phi,\pi,e^\alpha]$ vanishes, because we differentiate $\mathring{H}(x)$, first with respect to the multiplier λ^a, getting thus the λ equations. Therefore,

$$[*\mathring{H}(x),*\mathring{H}(x')]=[\mathring{H}(x),\ \mathring{H}(x)]_{\lambda^a=\lambda^a(\phi,\pi,e^\alpha)}, \tag{12.44}$$

with similar equations holding for any other pairs of constraint functions. The right-hand sides of these equations are given by the closing relations (12.31)–(12.33), in which all terms containing $[\partial\mathring{H}/\partial\lambda^a]_{\lambda^b=\lambda^b(\phi,\pi,e^\alpha)}$ are killed. This leads to the simplied closing relations between the modified constraint functions,

$$[*\mathring{H}(x),\ *\mathring{H}(x')]=-\epsilon(*\mathring{H}^a(x)\,\delta_{,a}(x,x')-(x\longleftrightarrow x')), \tag{12.45}$$

$$[{}^*\overset{\circ}{H}_a(x), {}^*\overset{\circ}{H}(x')] = {}^*\overset{\circ}{H}(x)\,\delta_{,a}(x, x'), \tag{12.46}$$

$$[{}^*\overset{\circ}{H}_a(x), {}^*\overset{\circ}{H}_b(x')] = \overset{\circ}{H}_b(x)\,\delta_{,a}(x, x') - (x \rightarrow\!\!\!\!\!- x'). \tag{12.47}$$

The closing relations (12.45)—(12.47) for a parametrized field theory on a flat Minkowskian background were first obtained by Dirac[3,4] who used a consistency argument to prove their validity. Here we have seen that the closing relations are the consequence of the foliation independence of the hypersurface action.

The closing relations (12.31)—(12.33) for scalar field, their generalization for an arbitrary tensor field, and the universal closing relations (12.45)—(12.47) for the constraint functions with eliminated λ multipliers, hold unchanged for the super-Hamiltonian $H(x)$ [or ${}^*H(x)$] and the supermomentum $H_a(x)$ [or ${}^*H_a(x)$] of the tensor fields interacting with the gravitational field. We will discuss them in the final paper of this series.

ACKNOWLEDGMENTS

I would like to thank Mr. James Nester for his comments on this paper.

*Work supported in part by the National Science Foundation under Grant No. GP-43718X to the University of Utah.

[1]K. Kuchař, J. Math. Phys. **17**, 777 (1976). Our notation is explained in Sec. 2 of that paper. The equations from it are quoted by prefixing the Roman numeral I before their section and equations numbers.

[2]K. Kuchař, J. Math. Phys. **17**, 792 (1976). The equations from it are quoted by prefixing the Roman numeral II before their section and equation numbers.

[3]See P.A.M. Dirac, *Lectures on Quantum Mechanics* (Academic, New York, 1965), and the papers quoted there.

[4]The parametrized scalar field is discussed by R. Arnowitt, S. Deser, and C.W. Misner, Phys. Rev. **116**, 1322 (1959). The constraints, however, are not projected into directions $\perp$ and $\parallel$ to the hypersurface. The electromagnetic field is discussed in R. Arnowitt, S. Deser, and C.W. Misner, The Dynamics of General Relativity, in: *Gravitation: An Introduction to Current Research*, edited by L. Witten (Wiley, New York, 1962). See also P.A.M. Dirac, Can. J. Math. **3**, 1 (1951).

[5]F. Belinfante, Phys. **7**, 449 (1940).

[6]L. Rosenfeld, Mem. Acad. R. Belg. Sci. **18**, No. 6 (1940).

[7]W. Pauli, Rev. Mod. Phys. **13**, 203 (1941).

[8]See, e.g., S.W. Hawking and G.F.R. Ellis, *The Large Scale Structure of Space—Time* (Cambridge U.P., Cambridge, 1973), p. 67.

[9]The use of the term "derivative gravitational coupling" is not entirely uniform. Sometimes, people refer by it to the theories in which the source and geometry are coupled through the terms containing the Riemann curvature tensor (the second derivatives of the metric).

[10]For the importance of the boundary terms in asymptotically flat spacetimes, see, e.g., B.S. DeWitt, Phys. Rev. **160**, 113 (1967), or T. Regge and C. Teitelboim, Ann. Phys. **88**, 286 (1974).

[11]K. Kuchař, Phys. Rev. D **4**, 955 (1971); K. Kuchař, J. Math. Phys. **13**, 768 (1972).

[12]K. Kuchař, "Canonical Quantization of Gravity," in *Relativity, Astrophysics and Cosmology*, edited by W. Israel (Reidel, Dordrecht 1973).

[13]C. Lanczos, *The Variational Principles of Mechanics* (University of Toronto Press, Toronto, 1970), 4th ed., Chaps. VI.10, IX.6—IX.10.

[14]F.B. Esfabrook and H.D. Wahlquist, SIAM Rev. **17**, 201 (1975).

Completely positive dynamical semigroups of *N*-level systems*

Vittorio Gorini† and Andrzej Kossakowski‡

Department of Physics, Center for Particle Theory, University of Texas at Austin, Austin, Texas 78712

E. C. G. Sudarshan

Department of Physics, Center for Particle Theory, University of Texas at Austin, Austin, Texas 78712 and Centre for Theoretical Studies, Indian Institute of Science, Bangalore 560012, India

(Received 19 March 1975)

We establish the general form of the generator of a completely positive dynamical semigroup of an N-level quantum system, and we apply the result to derive explicit inequalities among the physical parameters characterizing the Markovian evolution of a 2-level system.

I. INTRODUCTION

In this paper we establish the general form of the generator of a completely positive dynamical semigroup of an N-level quantum system (Sec. II) and we find the conditions, in the form of explicit inequalities, that complete positivity imposes on the physical parameters which characterize the Markovian evolution of a two-level system (Sec. III). The term *dynamical semigroup* was introduced by one of us to mean a continuous one parameter semigroup $\Lambda: t \to \Lambda_t$, $t \in \mathbb{R}^+$, of positive trace preserving linear maps Λ_t: $\mathcal{T}(\mathcal{H}) \to \mathcal{T}(\mathcal{H})$, where $\mathcal{T}(\mathcal{H})$ is the Banach space [under the trace norm $\|\sigma\|_1 = \mathrm{tr}(\sigma^*\sigma)^{1/2}$] of trace class operators on a complex separable Hilbert space $\mathcal{H}$.[1] Other terminologies which have been used in the literature are "quantum stochastic process"[2] and " (stationary) noncommutative Markov process."[3] Since Λ_t is a contraction,[4] it follows from the Hille-Yosida theorem[5] that there exists a linear operator $L: \mathcal{T}(\mathcal{H}) \to \mathcal{T}(\mathcal{H})$ with dense domain of definition $D(L)$ such that

$$\lim_{t \downarrow 0} \| L\sigma - t^{-1}(\Lambda_t \sigma - \sigma) \|_1 = 0, \quad \sigma \in D(L).$$

L is called the *generator* of the semigroup. Therefore, if we regard $\mathcal{H}$ as the Hilbert space associated to some quantum system, we can interpret Λ_t as the integrated form of a Markovian master equation for the density operator representing the state of the system

$$\frac{d\rho}{dt} = L\rho, \quad \rho \in \mathcal{T}(\mathcal{H}), \quad \rho \geq 0, \quad \mathrm{tr}(\rho) = 1. \tag{1.1}$$

Master equations of the form (1.1) are encountered in a wide variety of physical problems such as quantum optics, laser action, superradiance, oscillator damping, atomic and spin relaxation, decay of unstable systems, etc.[6–14] Generally speaking, an equation of the form (1.1) gives a correct description of the irreversible evolution of a quantum open system in contact with stationary surroundings, provided the decay time τ_R of the correlations of the "reservoir" is much shorter than the typical relaxation times τ_S of the system, so that memory effects can be neglected. If the latter condition is not met, one has in principle to solve for ρ a formally more complicated intergrodifferential equation with memory which is usually referred to as the generalized master equation (gme).[8,15–19] Recently, it has been shown by Davies that under suitable assumptions the gme does indeed go over into an equation of the form (1.1) with a rescaled time variable in the limit when the coupling of the system to its surroundings is made to tend to zero (weak-coupling limit, $\tau_S \to \infty$).[20] It is also possible to obtain (1.1) rigorously in the limit $\tau_R \to 0$. This has been called the limit of *singular reservoir*.[21] See our next paper for an explicit model thereof.[22]

In order to proceed further we need to recall the notion of completely positive map. Let $M(n)$ denote the C^* algebra of the $n \times n$ complex matrices and 1_n the identity map $M(n) \to M(n)$. A linear map $\alpha: \mathcal{A} \to \mathcal{B}$, $\mathcal{A}$ and $\mathcal{B}$ C^* algebras, is said to be *completely positive* if the tensor product map $\alpha^{(n)} = \alpha \otimes 1_n: \mathcal{A} \otimes M(n) \to \mathcal{B} \otimes M(n)$ is positive for all positive integers n (if $\alpha^{(p)}$ is positive for a given positive integer p, then α is called p positive).[23] For the theory of positive and completely positive maps of C^* algebras see Refs. 23–30. To show that complete positivity is actually a stronger condition than positivity, we give in Appendix A a general example of a positive map which is not two positive. Now let Λ be a dynamical semigroup and let $\mathcal{B}(\mathcal{H})$ denote the C^* algebra of bounded operators on $\mathcal{H}$. Let $\Lambda^*: t \to \Lambda_t^*$, $t \in \mathbb{R}^+$, be the positive, normal (i.e., ultraweakly continuous), and identity, preserving semigroup $\Lambda_t^*: \mathcal{B}(\mathcal{H}) \to \mathcal{B}(\mathcal{H})$, dual to Λ, defined by

$$\mathrm{tr}[(\Lambda_t \sigma)A] = \mathrm{tr}[\sigma(\Lambda_t^* A)], \quad \sigma \in \mathcal{T}(\mathcal{H}), \quad A \in \mathcal{B}(\mathcal{H}), \quad t \in \mathbb{R}^+ \tag{1.2}$$

(Λ^* provides the evolution in the Heisenberg picture). We say that Λ is a completely positive dynamical semigroup if the map Λ_t^*, $t \in \mathbb{R}^+$, is completely positive. One can argue that dynamical semigroups describing the evolution of physical systems should be completely positive. Indeed, assume we have a quantum system S coupled to a reservoir R. If we regard the total system $S+R$ as isolated, its dynamics will be given by a one-parameter group $U: t \to U_t$ of unitary transformations of $\mathcal{H}_S \otimes \mathcal{H}_R$, the tensor product of the Hilbert spaces associated to S and to R, respectively. Assume that $S+R$ has been initially prepared in a product state $\rho \otimes \sigma$, $\rho \in \mathcal{T}(\mathcal{H}_S)$, $\sigma \in \mathcal{T}(\mathcal{H}_R)$, in which S and R are uncorrelated. The Heisenberg reduced dynamics of S, $\Phi: t \to \Phi_t: \mathcal{B}(\mathcal{H}_S) \to \mathcal{B}(\mathcal{H}_S)$, $t \in \mathbb{R}^+$, is defined by

$$\mathrm{tr}\{(\rho \otimes \sigma)[U_t^*(A \otimes \mathbb{1})U_t]\} = \mathrm{tr}_S[\rho(\Phi_t A)], \tag{1.3}$$

where tr_S denotes the trace on $\mathcal{T}(\mathcal{H}_S)$. It is easy to see that Φ_t is completely positive. The proof can be found in a paper by Kraus,[25] who was the first, to our knowledge, to recognize the physical significance of complete positivity, in connection with state changes produced by quantum measurements. For the reader's convenience, we give in Appendix B a straightforward independent proof based on the definition. One can show by continuity that complete positivity will not be destroyed by any of the limiting procedures, such as weak-coupling or the singular-reservoir limit, which give rise from Φ_t to a dynamical semigroup. Other arguments justifying complete positivity have been given by Accardi[3] and by Lindblad,[31] which are based on the requirement of positivity, respectively, of quasiconditional expectations on the algebras of local (in time) observables and of the dynamics of the system $S+S'$, where S' is an auxiliary N-level system coupled trivially to the open system S. Both these arguments do not make reference to the dynamics of S being a subdynamics of a global unitary dynamics. We have received Lindblad's preprint after the completion of the first version of the present paper. In his work, using methods different from ours, the author gives the general form of the generator of a norm continuous completely positive dynamical semigroup. This result generalizes our theorem 2.2.

II. DYNAMICAL SEMIGROUPS OF N-LEVEL SYSTEMS

We now proceed to determine the structure of the generator of a completely positive dynamical semigroup of an N-level system. For such a system, we have the identifications $\mathcal{T}(\mathcal{H})=\mathcal{B}(\mathcal{H})=M(N)$ and if Λ is a completely positive dynamical semigroup thereof, it is clear that the map $\Lambda_t: M(N)\to M(N)$, $t\in\mathbb{R}^+$, is completely positive. We call Λ a completely positive dynamical semigroup of $M(N)$.

Let $\mathcal{P}_N$ denote the set of all complete families $\{P_1, P_2, \ldots, P_N\}$ of mutually orthogonal one-dimensional self-adjoint projections in $M(N)$: $P_iP_j=\delta_{ij}P_i$, $P_i=P_i^*$, $\sum_{l=1}^N P_l=1$. The following theorem is a special case of theorem 5 of Ref. 32:

Theorem 2.1. In order for a linear map $L: M(N)\to M(N)$ to be the generator of a dynamical semigroup of $M(N)$ it is necessary and sufficient that the conditions

$$\mathrm{tr}[P_r(LP_s)]\geq 0, \quad r\neq s=1,2,\ldots,N \tag{2.1}$$

and

$$\sum_{r=1}^N \mathrm{tr}[P_r(LP_s)]=0, \quad s=1,2,\ldots,N \tag{2.2}$$

hold for all $\{P_1, P_2, \ldots, P_N\}\in\mathcal{P}_N$. Condition (2.2) is necessary and sufficient for L to generate a trace preserving semigroup, whereas (2.1) expresses the positivity requirement.

Theorem 2.2. A linear operator $L: M(N)\to M(N)$ is the generator of a completely positive dynamical semigroup of $M(N)$ if it can be expressed in the form

$$L:\rho\to L\rho=-i[H,\rho]+\frac{1}{2}\sum_{i,j=1}^{N^2-1} c_{ij}\{[F_i,\rho F_j^*]+[F_i\rho,F_j^*]\}, \quad \rho\in M(N), \tag{2.3}$$

where $H=H^*$, $\mathrm{tr}(H)=0, \mathrm{tr}(F_i)=0$, $\mathrm{tr}(F_i^*F_j)=\delta_{ij}$, $(i,j=1,2,\ldots,N^2-1)$, and $\{c_{ij}\}$ is a complex positive matrix. For a given L, H is uniquely determined by the condition $\mathrm{tr}(H)=0$ and $\{c_{ij}\}$ is uniquely determined by the choice of the F_i's.

Remark. We may call $-i[H,\cdot]$ the "Hamiltonian" part of the generator and $L+i[H,\cdot]$ its dissipative part. In general, H is not the same as the Hamiltonian H_0 of the free N-level system.[20,22] The proof of theorem 2.2 is based on some lemmas.

Lemma 2.1 $t\to\Lambda_t$ is a completely positive dynamical semigroup of $M(N)$ iff $t\to\Lambda_t\otimes 1_N$ is a dynamical semigroup of $M(N)\otimes M(N)$.

Proof. From theorem 5 of Ref. 28, a linear map $\Gamma: M(N)\to M(N)$ is completely positive iff $\Gamma\otimes 1_N$ is positive. Expressing an element

$$\hat{A}\in M(N)\otimes M(N) \text{ as } \hat{A}=\sum_{i,j=1}^N A_{ij}\otimes E_{ij},\ A_{ij}\in M(N),$$

$(E_{ij})_{rs}=\delta_{ir}\delta_{js}$ and denoting by Tr the trace on $M(N)\otimes M(N)$ we have

$$\mathrm{Tr}[(\Gamma\otimes 1_N)\hat{A}]=\sum_{i,j=1}^N \mathrm{Tr}(\Gamma A_{ij}\otimes E_{ij})=\sum_{i,j=1}^N \mathrm{tr}(\Gamma A_{ij})\,\mathrm{tr}(E_{ij})=\sum_{i=1}^N \mathrm{tr}(\Gamma A_{ii}).$$

Hence Γ is trace preserving iff $\Gamma\otimes 1_N$ is trace preserving. QED

Lemma 2.2. Let Γ be a linear operator $M(N)\to M(N)$ and let $\{F_\alpha\}_{\alpha=1,2,\ldots,N^2}$ be a complete orthonormal set (c.o.s) in $M(N)$, viz., $(F_\alpha,F_\beta)=\mathrm{tr}(F_\alpha^*F_\beta)=\delta_{\alpha\beta}$. Then Γ can be uniquely written in the form

$$\Gamma: A\to\Gamma A=\sum_{\alpha,\beta=1}^{N^2} c_{\alpha\beta}F_\alpha AF_\beta^*, \quad A\in M(N). \tag{2.4}$$

Moreover, if $\Gamma A^*=(\Gamma A)^*$, then $c_{\alpha\beta}=\langle c_{\beta\alpha}\rangle$.

Proof. First note that

$$\sum_{\alpha=1}^{N^2} F_\alpha^*AF_\alpha=\mathbb{1}\,\mathrm{tr}(A), \forall A\in M(N). \tag{2.5}$$

Indeed, the left-hand side of (2.5) is invariant under a change of c.o.s. $F_\alpha\to E_\alpha$ and choosing $\{E_\alpha\}=\{E_{ij}\}$, we have

$$\sum_{i,j=1}^N E_{ij}^*AE_{ij}=\sum_{i,j=1}^N E_{ji}AE_{ij}=\left(\sum_{j=1}^N E_{jj}\right)\left(\sum_{i=1}^N A_{ii}\right)=\mathbb{1}\,\mathrm{tr}(A).$$

Now let $\mathcal{L}(M(N))$ denote the vector space of linear operators $M(N)\to M(N)$ and let $\{G_\alpha\}$ be a c.o.s. in $M(N)$. $\mathcal{L}(M(N))$ becomes a unitary space with the inner product

$$\langle\Gamma,\Phi\rangle=\sum_{\alpha=1}^{N^2}(\Gamma G_\alpha,\Phi G_\alpha)=\sum_{\alpha=1}^{N^2}\mathrm{tr}[(\Gamma G_\alpha)^*(\Phi G_\alpha)].$$

Define

$$\Gamma_{\alpha\beta}: A\to\Gamma_{\alpha\beta}A=F_\alpha AF_\beta^* \quad (\alpha,\beta=1,2,\ldots,N^2). \tag{2.6}$$

Then $\{\Gamma_{\alpha\beta}\}$ is a c.o.s. in $\mathcal{L}(M(N))$. Indeed, using (2.5) we have

$$\langle \Gamma_{\alpha\beta}, \Gamma_{\mu\nu}\rangle = \sum_{\lambda=1}^{N^2} \mathrm{tr}[(\Gamma_{\alpha\beta}G_\lambda)^*(\Gamma_{\mu\nu}G_\lambda)]$$
$$= \sum_{\lambda=1}^{N^2} \mathrm{tr}[(F_\alpha G_\lambda F_\beta^*)^*(F_\mu G_\lambda F_\nu^*)]$$
$$= \mathrm{tr}\left[F_\beta\left(\sum_{\lambda=1}^{N^2} G_\lambda^* F_\alpha^* F_\mu G_\lambda\right)F_\nu^*\right]$$
$$= \mathrm{tr}(F_\alpha^* F_\mu)\,\mathrm{tr}(F_\nu^* F_\beta) = \delta_{\alpha\beta}\delta_{\beta\nu}.$$

The last assertion of the lemma is now easily verified.

QED

Lemma 2.3. Let $\{F_\alpha\}_{\alpha=1,2,\ldots,N^2}$ be a c.o.s. in $M(N)$ such that $F_{N^2} = (1/N)^{1/2}\,\mathbb{1}$ and let L be a linear operator $M(N) \to M(N)$ such that $(LA)^* = LA^*$ and $\mathrm{tr}(LA) = 0$ for all $A \in M(N)$. Then L can be uniquely written in the form

$$L: A \to LA = -i[H, A] + \frac{1}{2}\sum_{i,j=1}^{N^2-1} c_{ij}\{[F_i, AF_j^*] + [F_iA, F_j^*]\}, \quad (2.7)$$

where $H = H^*$, $\mathrm{tr}(H) = 0$, and $c_{ij} = \langle c_{ji}\rangle_{\mathrm{av}}$.

Proof. From (2.4) we have

$$LA = \frac{1}{N}c_{N^2N^2}A + \left(\frac{1}{N}\right)^{1/2}\sum_{i=1}^{N^2-1}(c_{iN^2}F_iA + c_{N^2i}AF_i^*) + \sum_{i,j=1}^{N^2-1} c_{ij}F_iAF_j^* = -i[H,A] + \{G,A\} + \sum_{i,j=1}^{N^2-1} c_{ij}F_iAF_j^*, \quad (2.8)$$

where $H = (1/2i)(F^* - F)$ and $G = (1/2N)C_{N^2N^2}\mathbb{1} + (1/2)(F^* + F)$, with $F = (1/N)^{1/2}\sum_{i=1}^{N^2-1} c_{iN^2}F_i$. Now

$$0 = \mathrm{tr}(LA) = \mathrm{tr}\left[\left(2G + \sum_{i,j=1}^{N^2-1} c_{ij}F_j^*F_i\right)A\right], \quad \forall\ A \in M(N)$$

implies $G = -\frac{1}{2}\sum_{i,j=1}^{N^2-1} c_{ij}F_j^*F_i$, whence (2.7) follows. The uniqueness follows from dimensionality considerations, since $\mathrm{tr}(LA) = 0$, $\forall\ A \in M(N)$ implies N^2-independent conditions on L. QED

Lemma 2.4. Let $\{F_\alpha\}_{\alpha=1,2,\ldots,N^2}$ be a c.o.s. in $M(N)$. Then

$$\left\{\hat{P}_{(\alpha)} \;\middle|\; \hat{P} = \sum_{i,j=1}^{N} P_{ij}\otimes E_{ij};\ P_{ij} = F_\alpha E_{ij}F_\alpha^*;\ \alpha = 1,2,\ldots,N^2\right\}$$

is a complete family of mutually orthogonal self-adjoint projections in $M(N)\otimes M(N)$.

Proof. An element $\hat{P} = \sum_{i,j=1}^{N^2} P_{ij}\otimes E_{ij}$ of $M(N)\otimes M(N)$ is a self-adjoint projection iff

$$P_{ij}^* = P_{ji} \quad \text{and} \quad \sum_{l=1}^{N} P_{il}P_{lj} = P_{ij} \quad (i,j = 1,2,\ldots,N). \quad (2.9)$$

Two such projections $\hat{P}$ and $\hat{Q}$ are orthogonal iff

$$\sum_{l=1}^{N} P_{il}Q_{lj} = 0 \quad (i,j = 1,2,\ldots,N). \quad (2.10)$$

We have

$$P_{ij}^* = (F_\alpha E_{ij}F_\alpha^*)^* = F_\alpha E_{ij}^* F_\alpha^* = P_{ji}$$

and

$$\sum_{l=1}^{N} P_{il}P_{lj} = \sum_{l=1}^{N} F_\alpha E_{il}F_\alpha^* F_\beta E_{lj}F_\beta^* = F_\alpha E_{ij}F_\beta^*\,\mathrm{tr}(F_\alpha^*F_\beta) = \delta_{\alpha\beta}P_{ij}.$$

QED

Proof of theorem 2.2. The "if" part: If $t \to \Lambda_t$ is the semigroup generated by (2.3), the generator of the semigroup $t \to \Lambda_t \otimes 1_N$ is $L \otimes 1_N$. By Lemma 2.1 we must show that $\{c_{pq}\} \geq 1$ implies $L \otimes 1_N$ to satisfy the conditions of Theorem 2.1. Since $\mathrm{tr}(L\rho) = 0$ for all $\rho \in M(N)$, we need only check that

$$\mathrm{Tr}\left\{\hat{P}_{(1)}\left[(L\otimes 1_N)\,\hat{P}_{(2)}\right]\right\} \geq 0$$

for all pairs $\hat{P}_{(1)}$, $\hat{P}_{(2)}$ of mutually orthogonal self-adjoint projections in $M(N)\otimes M(N)$. And indeed, using (2.9) and (2.10), we get

$$\mathrm{Tr}\left\{\hat{P}_{(1)}\left[(L\otimes 1_N)\,\hat{P}_{(2)}\right]\right\} = \sum_{i,j=1}^{N}\mathrm{tr}\left[P_{ij}\left(L\,P_{ji}\right)\right]$$
$$= -i\sum_{i,j=1}^{N}\mathrm{tr}\left(P_{ij}\left[H, P_{ji}\right]\right) + \sum_{p,q=1}^{N^2-1} c_{pq}\sum_{i,j=1}^{N}\left[\mathrm{tr}\left(P_{ij}F_pP_{ji}F_q^*\right) - \tfrac{1}{2}\mathrm{tr}\left(P_{ij}F_q^*F_pP_{ji} + P_{ij}P_{ji}F_q^*F_p\right)\right]$$
$$= \sum_{p,q=1}^{N^2-1} c_{pq}\sum_{i,j=1}^{N}\mathrm{tr}\left(P_{ij}F_pP_{ji}F_q^*\right)$$
$$= \sum_{p,q=1}^{N^2-1} c_{pq}\sum_{i,j,k,l=1}^{N}\mathrm{tr}\left(P_{ik}P_{kj}F_pP_{jl}P_{li}F_q^*\right)$$
$$= \sum_{k,l=1}^{N}\sum_{p,q=1}^{N^2-1} c_{pq}\,\mathrm{tr}\left[\left(\sum_{j=1}^{N}P_{kj}F_pP_{jl}\right)\times\left(\sum_{j=1}^{N}P_{kj}F_qP_{jl}\right)^*\right] \geq 0,$$

since $\{c_{pq}\} \geq 0$.

The "only if" part: If a linear operator $L: M(N) \to M(N)$ generates a completely positive dynamical semigroup of $M(N)$ we have $\mathrm{tr}(LA) = 0$ and $(LA)^* = LA^*$ for all $A \in M(N)$. Hence, by Lemma 2.3, L can be written in the form (2.3) with $H = H^*$, $\mathrm{tr}(H) = 0$, and $c_{ij} = \langle c_{ji}\rangle_{\mathrm{av}}$. Since the matrix $\{c_{ij}\}$ is self-adjoint, we can choose another orthonormal set of traceless matrices $\{G_1, G_2, \ldots, G_{N^2-1}\}$ such that

$$L\rho = -i[H,\rho] + \frac{1}{2}\sum_{p=1}^{N^2-1}\lambda_p\{[G_p, \rho G_p^*] + [G_p\rho, G_p^*]\}, \quad \rho \in M(N).$$

Define

$$\hat{P}_{(q)} = \sum_{i,j=1}^{N}(G_qE_{ij}G_q^*)\otimes E_{ij}, \quad q = 1,2,\ldots,N^2-1 \quad \text{and}$$
$$\hat{P} = \frac{1}{N}\sum_{i,j=1}^{N} E_{ij}\otimes E_{ij}.$$

Then, by Lemma 2.4, Theorem 2.1, and Lemma 2.1 we have

$$0 \leq N\,\mathrm{Tr}\left\{\hat{P}_{(q)}[(L\otimes 1_N)\hat{P}]\right\}$$

$$= \sum_{p=1}^{N^2-1} \lambda_p \sum_{i,j=1}^{N} \operatorname{tr}(G_q E_{ij} G_q^* G_p E_{ji} G_p^*)$$

$$= \sum_{p=1}^{N^2-1} \lambda_p \operatorname{tr}(G_q^* G_p) \operatorname{tr}(G_q G_p^*) = \lambda_q, \quad q = 1, 2, \ldots, N^2 - 1.$$

The uniqueness of H and of $\{c_{ij}\}$ follows from Lemma 2.3. QED

III. TWO-LEVEL SYSTEM

In Ref. 33, Theorem 2.1 was applied to give the following characterization of the generator of a dynamical semigroup of $M(2)$.

Theorem 3.1. A linear operator $L: M(2) \to M(2)$ is the generator of a dynamical semigroup $t \to \Lambda_t$ of $M(2)$ iff it can be written in the form

$$L: \rho \to L\rho = -i[H, \rho] + \frac{1}{2} \sum_{i,j=1}^{3} c_{ij} \{[F_i, \rho F_j] + [F_i \rho, F_j]\}, \quad \rho \in M(2), \tag{3.1}$$

where (i) $H = \sum_{i=1}^{3} h_i F_i$, $h_i \in \mathbb{R}$;

(ii) $F_i = F_i^*$ and

$$F_i F_j = \tfrac{1}{4}\delta_{ij} \mathbb{1} + \frac{i}{2} \sum_{k=1}^{3} \epsilon_{ijk} F_k \left(\Rightarrow \operatorname{tr}(F_i F_j) = \tfrac{1}{2}\delta_{ij}, \ \operatorname{tr}(F_i) = 0\right);$$

(iii) $$\{c_{ij}\} = \begin{pmatrix} \gamma - 2\gamma_1 & -ia_3 & ia_2 \\ ia_3 & \gamma - 2\gamma_2 & -ia_1 \\ -ia_2 & ia_1 & \gamma - 2\gamma_3 \end{pmatrix}, \quad \gamma = \gamma_1 + \gamma_2 + \gamma_3;$$

(iv) $\gamma_1, \gamma_2, \gamma_3 \geq 0$;

(v) $a_i = \gamma_i m_i^0 + \sum_{j,k=1}^{3} \epsilon_{ijk} m_j^0 h_k$;

(vi) $m_i^0 = 0$ if $\gamma_1\gamma_2\gamma_3 = 0$;

(vii) $(m_1^0, m_2^0, m_3^0) \in S = \Big\{(z_1, z_2, z_3) \,\Big|\, z_1, z_2, z_3 \in \mathbb{R}$;

$$\inf_{x_1^2+x_2^2+x_3^2=1} \left[\sum_{i=1}^{3} \left(\gamma_i x_i (x_i - z_i) + \sum_{j,k=1}^{3} \epsilon_{ijk} x_i h_j z_k\right)\right] \geq 0; \ x_1, x_2, x_3 \in \mathbb{R} \Big\} \text{ if } \gamma_1\gamma_2\gamma_3 > 0.$$

Let $\rho_t = \Lambda_t \rho_0$ be the density matrix describing the system at time $t \geq 0$ and define the polarization components $M_i(t) = \operatorname{tr}(\rho_t F_i)$, $i = 1, 2, 3$. One easily verifies that the latter satisfy the following equations of motion (Bloch equations[34]):

$$\frac{dM_i(t)}{dt} = \sum_{j,k=1}^{3} \epsilon_{ijk} h_i (M_k(t) - M_k^0) - \gamma_i (M_i(t) - M_i^0), \quad i = 1, 2, 3, \tag{3.2}$$

where $M_i^0 = \frac{1}{2} m_i^0$ $(i = 1, 2, 3)$. $\mathbf{M}^0$ is a stationary state and it is the only stationary state iff $\gamma_1\gamma_2\gamma_3 > 0$ (in the latter case every state approaches $\mathbf{M}^0$ as $t \to \infty$).

If, for instance, we think of $M(2)$ as the algebra of observables of a spin-$\frac{1}{2}$ magnetic moment, we can interpret Eqs. (3.2) as describing spin relaxation in a molecular surrounding under the action of an external magnetic field $\mathbf{H} = (1/\hbar g)\mathbf{h}$, g being the gyromagnetic ratio. γ_1, γ_2, and γ_3 are damping factors which are directly related to the relaxation times of the polarization components towards their equilibrium values; they are in fact inverse relaxation times $\gamma_i = 1/T_i$, if L commutes with its Hamiltonian part $-i[H, \cdot\,]$.

If $\gamma = 0$, we have $L = -i[H, \cdot\,]$. This corresponds to a purely Hamiltonian evolution which is of course completely positive. Let $\gamma > 0$ and define $\kappa_i = \gamma - 2\gamma_i$. Then, it follows from Theorem 2.2 that in order for the evolution to be completely positive it is necessary and sufficient that

$$\begin{aligned} &\text{(a)} \ \kappa_1 + \kappa_2 + \kappa_3 \geq 0, \\ &\text{(b)} \ \kappa_2\kappa_3 + \kappa_3\kappa_1 + \kappa_1\kappa_2 \geq a_1^2 + a_2^2 + a_3^2, \\ &\text{(c)} \ \kappa_1\kappa_2\kappa_3 \geq \sum_{i=1}^{3} \kappa_i a_i^2. \end{aligned} \tag{3.3}$$

Conditions (3.3) are equivalent to the following:

$$\begin{aligned} &\text{(a)} \ \kappa_1, \kappa_2, \kappa_3 \geq 0, \\ &\text{(b)} \ a_1 = (\kappa_2\kappa_3)^{1/2} y_1, \\ &\text{(c)} \ a_2 = (\kappa_3\kappa_1)^{1/2} y_2, \\ &\text{(d)} \ a_3 = (\kappa_1\kappa_2)^{1/2} y_3, \\ &\text{(e)} \ y_1^2 + y_2^2 + y_3^2 \leq 1. \end{aligned} \tag{3.4}$$

In terms of the γ_i's, (3.4) (a) can be written

$$\gamma_1 + \gamma_2 \geq \gamma_3, \quad \gamma_2 + \gamma_3 \geq \gamma_1, \quad \gamma_3 + \gamma_1 \geq \gamma_2, \tag{3.5}$$

showing that no two relaxation times can be much longer than the third.

In particular, we see that no two γ_i's can be zero without the third being zero too. Hence a non-Hamiltonian completely positive evolution admits for at most a one dimensional manifold of equilibrium states. This is the case when one of the γ_i's, say γ_1, is zero. Then $\gamma_2 = \gamma_3$ and there is essentially only one relaxation time.

As a special example, we consider the case $\gamma_1 = \gamma_2 = \gamma_\perp > 0$, $\gamma_3 = \gamma_\parallel > 0$, and $h_1 = h_2 = 0$. Then we have $\kappa_1 = \kappa_2 = \gamma_\parallel$, $\kappa_3 = 2\gamma_\perp - \gamma_\parallel$ and conditions (3.4) become

$$\begin{aligned} &\text{(a)} \ 2\gamma_\perp \geq \gamma_\parallel, \\ &\text{(b)} \ a_i = [\gamma_\parallel(2\gamma_\perp - \gamma_\parallel)]^{1/2} y_i, \quad i = 1, 2, \quad a_3 = \gamma_\parallel y_3, \\ &\qquad y_1^2 + y_2^2 + y_3^2 \leq 1. \end{aligned} \tag{3.6}$$

For the equilibrium state we get

$$M_1^0 = [\gamma_\parallel(2\gamma_\perp - \gamma_\parallel)]^{1/2} \left(\frac{\gamma_\perp y_1 - h_3 y_2}{2(\gamma_\perp^2 + h_3^2)}\right),$$

$$M_2^0 = [\gamma_\parallel(2\gamma_\perp - \gamma_\parallel)]^{1/2} \left(\frac{\gamma_\perp y_2 + h_3 y_1}{2(\gamma_\perp^2 + h_3^2)}\right), \quad M_3^0 = y_3. \tag{3.7}$$

If the system is rotationally symmetric about the direction of the magnetic field, we have $M_1^0 = M_2^0 = 0$. In this case $\gamma_\perp$ and $\gamma_\parallel$ are, respectively, the inverse transverse and the inverse longitudinal relaxation times and (3.6) (a) is written

$$T_\parallel \geq \tfrac{1}{2} T_\perp, \tag{3.8}$$

a relation which had been previously derived by Favre and Martin for a spin system weakly coupled to a high-temperature bath.[18] To our knowledge, relation (3.8) is experimentally satisfied in all known cases.

ACKNOWLEDGMENTS

We thank the referee for useful comments.

APPENDIX A

The following proposition provides a fairly general example of a positive map which is not two positive.

Proposition. Let $\mathcal{A}$ be a non commutative C* algebra which identity and let β be a $*$ antiautomorphism of $\mathcal{A}$. Then β is not two positive.

Proof. Let $\hat{A}$ be a self-adjoint element of $\mathcal{A}\otimes M(2)$. It has the form $\hat{A}=\sum_{i,j=1}^{2}A_{ij}\otimes E_{ij}$, where $(E_{ij})_{rl}=\delta_{ir}\delta_{jl}$ and $A_{ij}^{*}=A_{ji}$, and we have

$$\beta^{(2)}(\hat{A}^2)-[\beta^{(2)}(\hat{A})]^2=\beta[A_{12},A_{12}^{*}]\otimes E_{11}+\beta[A_{11}-A_{22},A_{12}]\otimes E_{12}-\beta[A_{11}-A_{22},A_{12}^{*}]\otimes E_{21}-\beta[A_{12},A_{12}^{*}]\otimes E_{22}.$$

Assume $\beta^{(2)}$ is positive. Then, since $\beta^{(2)}$ is self-adjoint and identity preserving, we have $\|\beta^{(1)}\|=1$.[35] Therefore, $\beta^{(2)}$ satisfies Kadison's inequality[36] $\beta^{(2)}(\hat{A}^2)-[\beta^{(2)}(\hat{A})]^2\geq 0$. By the above, this implies $\beta[A_{12},A_{12}^{*}]=0$, which, by the arbitrariness of A_{12}, contradicts the noncommutativeness of $\mathcal{A}$. QED

APPENDIX B

To simplify notations we drop the subscript t from Φ_t and U_t and write $\rho(A)$ in place of $\mathrm{tr}(\rho A)$. An element $\hat{A}\in\mathcal{B}(\mathcal{H}_S)\otimes M(n)$ admits of a unique decomposition $\hat{A}=\sum_{i,l=1}^{n}A_{il}\otimes E_{il}$, $A_{il}\in\mathcal{B}(\mathcal{H}_S)$, $(E_{il})_{rs}=\delta_{ir}\delta_{ls}$, and if $\hat{A}$ is positive we have

$$\hat{A}=\hat{B}^{*}\hat{B}=\sum_{i,j=1}^{n}\left(\sum_{l=1}^{n}B_{li}^{*}B_{lj}\right)\otimes E_{ij}.$$

A density operator on $\mathcal{B}(\mathcal{H}_S)\otimes M(n)$ can be written as a finite convex combination of states of the form $\rho\otimes\omega$, where ρ is a density operator on $\mathcal{B}(\mathcal{H}_S)$ and ω is a pure state on $M(n)$, viz., $\omega(E_{ij})=\bar{x}_i x_j$. Set $Q_{ki}=U^{*}(B_{ki}\otimes 1)U$ and $Q_s=\sum_{r=1}^{n}x_r Q_{sr}$. Then

$$\begin{aligned}(\rho\otimes\omega)[\Phi^{(n)}(\hat{B}^{*}\hat{B})]&=(\rho\otimes\omega)\left[\sum_{i,j}\left(\sum_{k}\Phi(B_{ki}^{*}B_{kj})\otimes E_{ij}\right)\right]\\&=\sum_{i,j,k}\rho[\Phi(B_{ki}^{*}B_{kj})]\bar{x}_i x_j\\&=\sum_{i,j,k}(\rho\otimes\sigma)[U^{*}(B_{ki}^{*}B_{kj}\otimes 1)U]\bar{x}_i x_j\\&=\sum_{i,j,k}\bar{x}_i(\rho\otimes\sigma)(Q_{ki}^{*}Q_{kj})x_j\\&=(\rho\otimes\sigma)\left(\sum_{s}Q_s^{*}Q_s\right)\geq 0.\end{aligned}$$ QED

*Supported in part by USAEC, Contract No. AT(40-1)3992, by INFN, Sezione di Milano and by the Institute of Mathematics, Polish Academy of Science, Warsaw.

†Permanent address: Istituto di Fisica dell'Università, via Celoria 16, 20133 Milano, Italy and INFN Sezione di Milano.

‡Permanent address: Institute of Physics, N. Copernicus University 87100 Toruń, Poland.

[1] A. Kossakowski, Rep. Math. Phys. **3**, 247 (1972).
[2] E. B. Davies, Commun. Math. Phys. **15**, 277 (1969); **19**, 83 (1970); **22**, 51 (1971).
[3] L. Accardi, "Non-Relativistic Quantum Mechanics as a Non-Commutative Markov Process," preprint Lab. di Cibernetica del C.N.R., Arco Felice (Naples), 1975.
[4] This property follows from Theorem 3.3 of Ref. 26 applied to the dual map Λ_t^{*} defined by formula (1.2).
[5] K. Yosida, *Functional Analysis* (Springer-Verlag, Berlin, 1972), 3rd ed.
[6] G. S. Agarwal, in *Progress in Optics*, edited by E. Wolf (North-Holland, Amsterdam, 1973), Vol. XI, pp. 1—75.
[7] H. Haken, *Laser Theory*, *Handbook of Physis*, 25/2c (Springer-Verlag, Berlin, 1970).
[8] F. Haake, in *Springer Tracts in Modern Physics*, Vol. 66, (Springer-Verlag, Berlin, 1973), pp. 98—168.
[9] W. Happer, Rev. Mod. Phys. **44**, 169 (1972).
[10] F. Bloch and R. K. Wangsness, Phys. Rev. **89**, 728 (1953).
[11] A. Abragam, *The Principles of Nuclear Magnetism* (Oxford U. P., London, 1961).
[12] N. M. Atherton, *Electron Spin Resonance* (Wiley, New York, 1973).
[13] S. Twareque Ali, L. Fonda, and G. C. Ghirardi, Nuovo Cimento A **25**, 134 (1974).
[14] A. Barchielli and L. Lanz, "Banach Space Representations of the Galilei Semigroup: Decay of an Unstable Particle," preprint, University of Milan, 1975.
[15] R. Zwanzig, J. Chem. Phys. **33**, 1338 (1960); and in *Lectures in Theoretical Physics (Boulder)*, edited by W. E. Brittin, B. W. Downs, and J. Downs (Interscience, New York, 1961), Vol. 3, p. 106.
[16] P. N. Argyres and P. L. Kelley, Phys. Rev. **134**, A98 (1964).
[17] G. G. Emch and G. L. Sewell, J. Math. Phys. **9**, 946 (1968).
[18] C. Favre and Ph. Martin, Helv. Phys. Acta **41**, 333 (1968).
[19] E. Presutti, E. Scacciatelli, G. L. Sewell, and F. Wanderlingh, J. Math. Phys. **13**, 1085 (1972).
[20] E. B. Davies, Commun. Math. Phys. **39**, 91 (1974). See also J. V. Pulé, Commun. Math. Phys. 38, 241 (1974), for a model which does not start directly from a gme.
[21] K. Hepp and E. H. Lieb, Helv. Phys. Acta **46**, 573 (1973).
[22] V. Gorini and A. Kossakowski, J. Math. Phys. (to be published). See also A. Frigerio and V. Gorini, J. Math. Phys. (to be published), and E. B. Davies, Z. Wahrscheinlichkeitstheorie Verw. Geb. 23, 261 (1972).
[23] W. F. Stingespring, Proc. Am. Math. Soc. **6**, 211 (1958).
[24] E. Størmer, Acta Math. **110**, 223 (1963).
[25] K. Kraus, Ann. Phys. (N.Y.) **64**, 311 (1971).
[26] E. Størmer, in *Lecture Notes in Physics*, edited by A. Hartkämper and H. Neumann (Springer-Verlag, Berlin, 1974), Vol. 29, pp. 85—106.
[27] W. Arveson, Acta Math. **123**, 141 (1969).
[28] M. D. Choi, Can. J. Math. **24**, 520 (1972).
[29] M. D. Choi, Linear Algebra Appl. (to be published).
[30] M. D. Choi, Ill. J. Math. **18**, 565 (1974).
[31] G. Lindblad, "On the Generators of Quantum Dynamical Semigroups," preprint Royal Institute of Technology, Stockholm, TRITA-TFY-75-1, 1975.
[32] A. Kossakowski, Bull. Acad. Pol. Sci. Sér. Math. Astr. Phys. **20**, 1021 (1972).
[33] A. Kossakowski, Bull. Acad. Pol. Sci. Sér. Math. Astr. Phys. **21**, 649 (1973).
[34] F. Bloch, Phys. Rev. **70**, 460 (1946).
[35] B. Russo and H. A. Dye, Duke Math. J. **33**, 413 (1966).
[36] R. V. Kadison, Ann. Math. **56**, 494 (1952).

A characteristic glimpse of the renormalization group

John R. Klauder

Bell Laboratories, Murray Hill, New Jersey 07974
(Received 15 September 1975)

Characteristic functions for arbitrary physical systems are discussed in general terms with special emphasis on the space of allowed test variables for which the characteristic function is defined. A fundamental change in the probability measure (at a phase boundary, for example) is reflected in a change of allowed test variables, generally, some new ones added and some old ones dropped. Physically motivated transformations on the space of test variables (implicitly introducing block spins, for example) are designed, after repeated operation, to accentuate or even to isolate the fundamental changes. Several examples of such renormalization-group type transformations are given.

Renormalization group ideas and methods have become all pervasive in the last few years and justifiably so.[1] The language of probability theory enables one to unify many of these ideas and partially relate them to certain limit theorems known in mathematics as has been emphasized recently by Jona-Lasinio.[2] Here we restate and extend some of these ideas in the language of characteristic functions which often provides a useful complementary description.

Quite generally, a characteristic function $C(h)$ is defined for a variety of test variables h and has the general structure

$$C(h)=\int \exp[i(h,\Phi)]\,d\mu(\Phi),$$

where h is a real variable (number, sequence, function, etc.), Φ denotes a generalized real random variable (number, sequence, function, generalized function, etc., discrete, continuous or both), μ is a probability measure on configurations (normalized Gibbs distribution, etc.), and (h,Φ) denotes the relevant (real) inner product. We are particularly interested in infinite-dimensional test variable spaces. Both statistical mechanics and certain Euclidean field theory problems are covered in the present generality, and a few examples will be given below.

The measure μ depends on the parameters of the problem (temperature, spin coupling, coefficients of nonlinearities, masses, cutoffs, etc.), and as a consequence so does the characteristic function C; in fact, given either μ or C the other quantity is uniquely determined. In general cases the support of μ is a complicated issue and we will not discuss it.[3] A complementary question, in some sense, is the *allowed class of test variables* and this is somewhat easier. In the general case, we note without proof that

$$d_C^2(h)\equiv \pi^{-1/2}\int [1-\mathrm{Re}C(\lambda h)]\exp(-\lambda^2)\,d\lambda$$
$$=\int \{1-\exp[-\tfrac{1}{4}(h,\Phi)^2]\}\,d\mu(\Phi)$$

defines a *metric* $d_C(h)$ on the space of test variables, that this space may be *completed* to include limits of Cauchy sequences, and that each element of the completed (linear vector) space $\mathcal{V}_C$ defines an *acceptable*.[4] *test variable*.[4] By construction, $C(h)$ is a *continuous function* on $\mathcal{V}_C$ in the topology induced by d_C. Commonly one starts with a *restricted and conservative* class of test variables, applicable to a wide variety of problems; however, the specific problem at hand, namely μ or C, ultimately determines the *maximally allowed* class of test variables, and this is as it should be. For instance, when a physical system crosses a phase boundary the support of μ generally changes fundamentally (e.g., revised long-range order, modified symmetry breaking, etc.), and there is a corresponding fundamental change in the characteristic function C that reflects itself in a fundamental change in the space of allowed test variables $\mathcal{V}_C$.

The renormalization group attempts to isolate the fundamental changes in μ, C, and $\mathcal{V}_C$ by selectively emphasizing one or another of the fundamental changes. To this end consider a general transformation T (real, linear, nonsingular) that maps the space of test variables into itself, i.e., $T:\mathcal{V}_C\to\mathcal{V}_C$. For the test variables themselves, $h\mapsto h_T\equiv Th$ for all allowed h. Such a transformation induces a transformation of C (and therefore of μ) according to

$$C(h)\mapsto C_T(h)\equiv C(Th).$$

Being a normalized, continuous, positive-definite function, $C_T(h)$ necessarily has the representation

$$C_T(h)=\int \exp[i(h,\Phi)]\,d\mu_T(\Phi),$$

which implicitly defines the map of μ to μ_T (generally, μ and μ_T are inequivalent measures). More explicitly we also note that

$$C_T(h)=\int \exp[i(Th,\Phi)]\,d\mu(\Phi)$$
$$=\int \exp[i(h,T^{\sim}\Phi)]\,d\mu(\Phi),$$

where $T^{\sim}$ denotes the transposed transformation. Although we choose T as nonsingular, it may well happen that $T^{\sim}$ is singular.

For the sake of illustration suppose Φ denotes a sequence of spin variables $\{\Phi_k;\ k=1,2,\cdots\}$, h is also a sequence $\{h_k;\ k=1,2,\cdots\}$, and $(h,\Phi)=\sum_{k=1}^{\infty}h_k\Phi_k$. Then T is a matrix $\{T_{kl}\}$, and we specifically choose $T_{kl}=p\delta_{[1+(k-1)/m],\,l}$, where $[x]$ denotes the largest integer in x, m is an integer, $m\geq 2$, and p is a parameter to be chosen as needed. Consequently,

$$(Th,\Phi)=(h,T^{\sim}\Phi)=\sum_{k=1}^{\infty}h_k\Big(p\sum_{l=1+m(k-1)}^{mk}\Phi_l\Big),$$

which leads directly to a block spin encompassing m "old" spins in each "new" spin. In this example, therefore, μ_T, as determined by C_T, is the probability measure for block spins. Observe that a transformation on test variables has in effect introduced a conjugate transformation of physical interest on the spin variables.

Any given transformation T may be repeated arbitrarily often which leads to the sequence of characteristic functions

$$C^{(n)}(h) \equiv C(T^n h),$$

and one may ask whether

$$C^{(n)}(h) \to C^{(T)}(h),$$

as $n \to \infty$, where $C^{(T)}$ denotes a characteristic function. Evidently, if this is the case,

$$C^{(T)}(Th) = C^{(T)}(h),$$

and thus $C^{(T)}$ is invariant under T (or in other words, we have a fixed point of the transformation T).

Generally, however, a sequence of characteristic functions does not converge; more typically, there may exist one or more convergent *subsequences*. One sufficient condition to ensure convergent subsequences requires a metric $d_{C,T}$ (equivalent to d_C) for $\not{V}_C$ such that $d_C(T^n h) \leq d_{C,T}(h)$ uniformly in n. In this case there exists a subsequence $\{n_r; r = 1, 2, \cdots\}$ such that as $r \to \infty$ (and $n_r \to \infty$),

$$C^{(n_r)}(h) = C(T^{n_r} h) \to C^{(T)}(h).$$

The resultant characteristic function $C^{(T)}(h)$ is generally not invariant under T. For instance, in terms of test variables $\{h_k;\ k = \cdots, -2, -1, 0, 1, \cdots\}$ and $(h_T)_k \equiv h_{k+1}$, suppose the limiting function is

$$C^{(T)}(h) = \exp(-\sum a_k h_k^2).$$

If $n_r = r$, then a_k is required to be independent of k and $C^{(T)}$ is invariant under T. However, if $n_r = 2r$, then a_k is only restricted to be periodic with period 2, i.e., $a_k = f(k/2, \text{ mod } 1)$, for some $f(0)$ and $f(\frac{1}{2})$, and $C^{(T)}$ is invariant under T^2 but generally not under T. If, instead, $n_r = 2^r$, then a_k is restricted to be log periodic, i.e., for $k > 0$, $a_{\pm k} = f_\pm(\ln|k|/\ln 2, \text{ mod } 1)$, for some $f_\pm(x)$, $0 \leq x < 1$ and a_0 arbitrary. In the latter case, $C^{(T)}$ is generally noninvariant under T^m for *any* m.

Since a general transformation T leads only to a convergent subsequence of $C^{(n)}(h) = C(T^n h)$ interest centers on those special transformations for which $C^{(n)}(h) \to C^{(T)}(h)$. When a phase boundary is reached or crossed and special features change, the space of test variables can change in essentially just two ways: either by adding *new* vectors or deleting *old* vectors (or a combination of the two). The role of the renormalization group is to bring such changes into prominence. In the first type, the special feature is already revealed by special vectors in $\not{V}_C$, and in this case it suffices to arrange that $h^{(n)} \equiv T^n h$ is a Cauchy sequence in $\not{V}_C$, or more generally that $T^n : \not{V}_C$ is, in the limit, a map onto a subspace of $\not{V}_C$. Evidently the resultant subspace is invariant under T. In the second type, the special feature appears implicitly since the space $\not{V}_C$ has in part diminished. The effect of the requisite transformation is to expose the special feature, and this is generally accompanied by a change of the space of test variables. Specifically, if $d_C^{(n)}(h) \equiv d_C(T^n h) \leq d_C(h)$ converges to a metric $d_{C^{(T)}}(h)$, then the new space $\not{V}_{C^{(T)}}(h)$ is just the completion of the space $\not{V}_C(h)$ with respect to the metric $d_{C^{(T)}}(h)$. Correspondingly, $C^{(n)}(h) \to C^{(T)}(h)$, and both $C^{(T)}$ and $d_{C^{(T)}}$ are invariant under T. [It should perhaps be remarked that in the general case convergence of $d_C^{(n)}$ implies convergence of $C^{(n)}$ but not conversely. It is useful to note that convergence of $d_C^{(n)}(h)$ is equivalent to convergence of $C^{(n)}(\lambda h)$ for all real λ, and this often proves a fairly simple way to determine $\not{V}_C$.]

Some examples will illustrate the two types of behavior. First consider the simple case for which $C_0(h) = \exp(-a\sum h_k^2)$, where $a > 0$. The metric d_{C_0} is equivalent to that induced by the standard norm and $\not{V}_{C_0} = l^2 \equiv \{h_k : \sum h_k^2 < \infty\}$ independently of the variable a. Next consider

$$C(h) = \exp[-a\sum h_k^2 - b(\sum h_k)^2],$$

where $a, b > 0$, for which the space of allowed test variables is given by

$$\not{V}_C \equiv \{h_k : \sum h_k^2 < \infty, (\sum h_k)^2 < \infty\}.$$

(Schematically, the former expression is meant to apply above a critical temperature, the latter below.[5]) The spaces $\not{V}_{C_0}$ and $\not{V}_C$ are fundamentally different, and to illustrate this difference consider the transformation T specified previously (with $p = m^{-1}$) and choose $h_k = g_k \equiv \delta_{k1}$. Then for this specific test variable,

$$g_k^{(n)} = (T^n g)_k = m^{-n}, \quad 1 \leq k \leq m^n,$$
$$= 0, \qquad\qquad k > m^n.$$

With $b = 0$, $g^{(n)}$ is a Cauchy sequence in $\not{V}_{C_0} = l^2$ converging to *zero*, $g^{(n)} \to 0$; with $b > 0$, $g^{(n)}$ is a Cauchy sequence in $\not{V}_C$ converging to a *nonzero* element, $g^{(n)} \to \bar{g} \neq 0$, an element not contained at all in $\not{V}_{C_0} = l^2$. Stated otherwise, when $b = 0$, there is, as usual, an equivalence class of zero elements in l^2; effectively speaking, when $b > 0$, this equivalence class is "broken up" into some zero and some new nonzero elements. It is such novel elements of $\not{V}_C$ that the renormalization group attempts to bring into prominence. Specifically, with T as just defined,

$$C(T^n h) \to C^{(T)}(h) = \exp[-b(\sum h_k)^2],$$

which no longer depends on the variable a (a reflection of the concept of universality). Observe also that only *one* giant block spin survives in the limit here since

$$\exp[-b(\sum h_k)^2] = (4\pi b)^{-1/2} \int \exp[i(\sum h_k)S - \tfrac{1}{4}b^{-1}S^2]\, dS.$$

Moreover, we must also remark, since $T^n h$ is a Cauchy sequence our limiting function in this case amounts to no more than looking at a special subspace of $\not{V}_C$. In particular, for any $h \in \not{V}_C$, $T^n h \to h_1 \bar{g}$, where $\bar{g}$ is the specific element of $\not{V}_C$ introduced above. As $n \to \infty$, then, $T^n \not{V}_C$ collapses to a *one*-dimensional space, a subspace within $\not{V}_C$, with a basis vector $\bar{g}$. The vector $\bar{g}$ is evidently invariant under T. In summary, the study of the *limiting* fixed-point characteristic function

$$\hat{C}(y) \equiv C^{(T)}(h) = \exp(-by^2),$$

where $y \equiv \sum h_k$, is neither more nor less than the study of the *original* characteristic function for restricted test variables of the form $h = y\bar{g}$ since

$$C(y\bar{g}) = \exp(-by^2) = \hat{C}(y).$$

Of course, a given $C(h)$ may admit interesting limit points for several different transformations. One need only augment the previous example so that

$$C(h) = \exp\{-a\sum h_k^2 - b(\sum h_k)^2 - b'[\sum (-1)^k h_k]^2\}.$$

In this case we have in addition a "staggered magnetization," which is exposed by a transformation T' determined from T by

$$T'_{kl} = (-1)^k T_{kl}.$$

In particular, it follows that

$$C(T'^n h) \to C^{(T')}(h) = \exp\{-b'[\sum (-1)^k h_k]^2\}.$$

Just as in the preceding case, there is a one-dimensional subspace of $\mathcal{V}_C$ with basis vector $\bar{g}' = \lim T'^n g$ that provides the same information since

$$C(y'\bar{g}') = \exp(-b'y'^2).$$

Evidently, additional variations on this theme may be readily played.

We continue with another simple example appropriately described on a continuous space (chosen for convenience as three dimensional). Thus we consider $h(x)$, $x \in R^3$, and assume

$$C(h) = \exp[-a\int h^2(x)\,dx - b\int h(x)W(x-y)h(y)\,dx\,dy].$$

{When $a > 0$ and $b = 0$ [and $\mathcal{V}_C = L^2(R^3)$], we suppose we are above a critical temperature; when $a, b > 0$ we are at or below as we shall discuss.} Here $W(x)$ admits the general representation

$$W(x) = (2\pi)^{-3}\int \exp(ikx)\,d\sigma(k),$$

where σ is a positive measure. Consequently, we may also write

$$C(h) = \exp[-a\int |\tilde{h}(k)|^2\,dk - b\int |\tilde{h}(k)|^2\,d\sigma(k)].$$

Fixed point properties of this example depend strongly on the detailed nature of σ.

If $\sigma(k)$ is *discrete*, then

$$C(h) = \exp[-a\int |\tilde{h}(k)|^2\,dk - b\textstyle\sum_r \sigma_r |\tilde{h}(k_r)|^2],$$

where $\sigma_r > 0$ and k_r belong to the atoms of σ. In this case

$$\mathcal{V}_C = \{h : \int |\tilde{h}(k)|^2\,dk < \infty,\ |\tilde{h}(k_r)|^2 < \infty \text{ all } r\},$$

which contains elements not in L^2. Transformations T_r, where

$$h_{T_r}(x) = -(2L)^{-3}\exp(-ik_r x)\int_{x-L}^{x+L} h(x')\,dx'$$

for some $L > 0$, lead to the behavior

$$C(T_r^n h) \to C^{(T_r)}(h) = \exp[-b\sigma_r |\tilde{h}(k_r)|^2];$$

these cases are just equivalent to examining certain subspaces of $\mathcal{V}_C$ (exactly as in the discrete spin case below the critical temperature).

If $\sigma(k)$ is *absolutely continuous* the analysis is different. Let $d\sigma(k) = \tilde{W}(k)\,dk$, $\tilde{W}(k) \geq 0$, and then

$$\mathcal{V}_C = \{h : \int [1 + \tilde{W}(k)]|\tilde{h}(k)|^2\,dk < \infty\}.$$

If $\tilde{W}(k) \leq M$, then $\mathcal{V}_C = L^2(R^3)$ and no anomaly exists; but if $\tilde{W}(k)$ is *not* uniformly bounded $\mathcal{V}_C \subset L^2(R^3)$, and this happens generally at the critical temperature for some ordering. Commonly one considers as $|k| \to 0$ that $\tilde{W}(k) \to |k|^{-\sigma}$, $0 < \sigma \leq 2$ (but, of course, an unbounded behavior is not limited to $k = 0$). A scaling limit can be taken to isolate this long-range order by taking T to be defined by $h_T(x) = S^{d-3}h(S^{-1}x)$ or equivalently by $\tilde{h}_T(k) = S^d\tilde{h}(Sk)$, where $1 < S < \infty$ (S is the analog of m in the discrete case), and d will be fixed later. (The parameter d is a scale dimension and not a space dimension.) It follows that

$$\begin{aligned} C(Th) &= \exp\{-\int [a + b\tilde{W}(k)]|\tilde{h}_T(k)|^2\,dk\} \\ &= \exp\{-\int [a + b\tilde{W}(k)]S^{2d}|\tilde{h}(Sk)|^2\,dk\} \\ &= \exp\{-S^{2d-3}\int [a + b\tilde{W}(S^{-1}k)]|\tilde{h}(k)|^2\,dk\}. \end{aligned}$$

Under repeated transformations and as $n \to \infty$, only the small $|k|$ dependence of $\tilde{W}(k)$ is tested [or the large $|x|$ dependence of $W(x)$]. Since we assume that

$$\tilde{W}(k) \to |k|^{-\sigma}, \quad 0 < \sigma \leq 2,$$

choose $d = \frac{1}{2}(3 - \sigma)$, then only the parameter b survives and in fact

$$C^{(T)}(h) = \exp[-b\int |k|^{-\sigma}|\tilde{h}(k)|^2\,dk],$$

which is invariant under T. The long-range correlation here is given by $|x|^{-(3-\sigma)}$, which means that the standard critical index $\eta = 2 - \sigma$.[6]

From the point of view of metric convergence in this case, it is clear that the metric $d_C(h)$ is equivalent to the norm $\|h\|$ where

$$\|h\|^2 \equiv \int [1 + \tilde{W}(k)]|\tilde{h}(k)|^2\,dk.$$

Convergence of $d_C(T^n h)$ is equivalent to convergence of $\|T^n h\|$, which we have essentially computed since

$$\|T^n h\|^2 \to \|h\|^2_{(T)} \equiv \int |k|^{-\sigma}|\tilde{h}(k)|^2\,dk.$$

It is also clear that the completed space $\mathcal{V}_C^{(T)}$ contains elements that did not lie in the original $\mathcal{V}_C$.

As regards the approach to critical behavior as $t \to 0$, where $t \equiv (T - T_c)/T_c$, we observe for $\sigma = 2$ that short-range forces are involved while for $0 < \sigma < 2$ we effectively deal with a problem involving long-range forces. Typically one assumes[7] for $t > 0$ that

$$\tilde{W}(k) = r + |k|^\sigma + o(|k|^\sigma),$$

and that in natural units $r = t$ (analyticity assumption). Since for small k the relevant variable in $\tilde{W}(k)$ is $k/r^{1/\sigma}$, it follows for large x that the relevant variable in $W(x)$ is $xr^{1/\sigma}$ and thus the correlation length $\xi \sim r^{-1/\sigma} \approx K(T - T_c)^{-1/\sigma}$. The correlation length critical exponent ν is thus given by $\nu = 1/\sigma$. For short range forces $\sigma = 2$, and $\nu = \frac{1}{2}$; when suitable long range forces are present $\sigma < 2$, and $\nu > \frac{1}{2}$. While Gaussian systems are pedagogically helpful, interesting systems are generally non-Gaussian.

Simple and physically relevant non-Gaussian examples are hard to come by. If for convenience one accepts characteristic functions as "primary" rather than Hamiltonians (say) as "primary," one may readily propose non-Gaussian examples the physical signi-

ficance of which unfortunately remains rather obscure. As one easily analyzed non-Gaussian example, consider

$$C(h)=\exp\{-b\int |k|^{-3}dk\int [1-J_0(\lambda|k|^d|\tilde{h}(k)|)]\rho(\lambda)\,d\lambda\},$$

where $\rho(\lambda)\geq 0$. As presented this example is already invariant under the transformation $\tilde{h}_T(k)=S^d\tilde{h}(Sk)$, and [assuming $\rho(\lambda)$ has a second moment] the long-range behavior of the second-order correlation is characterized by the Fourier transform of $|k|^{2d-3}$, namely by $|x|^{-2d}$, which leads to a critical index $\eta=2d-1$. However, we can make various proposals away from the critical behavior, one such being to replace the term $|k|$ by $(r+|k|^\sigma)^{1/\sigma}$, $0<\sigma\leq 2$, where $r=t$, which has the consequence, just as in the Gaussian case, that $\xi\approx K(T-T_c)^{-1/\sigma}$. In this simple example we have illustrated a consistent model with two independently specifiable critical indices, ν and η. A conventional argument indicates that all other standard indices may be determined algebraically from these two.[6]

It is interesting to observe that the space of allowed test functions is not at all that suggested by the quadratic terms but rather is a larger space. To make this more evident assume $\rho(\lambda)=\exp(-\lambda)$, $\lambda>0$, so that

$$C(h)=\exp(-b\int |k|^{-3}dk\{1-[1+|k|^{2d}|\tilde{h}(k)|^2]^{-1/2}\}).$$

Then it turns out that singularities as strong as $\tilde{h}(k)=|k-\kappa|^{-p}$, $\kappa\neq 0$, are acceptable for *any* $p>0$. On the other hand, the quadratic component requires $\tilde{h}(k)$ locally square integrable for $k\neq 0$, and hence $p<\frac{3}{2}$. Such unusual behavior as exhibited here (and in the following example) is suggested by certain model field theories.[8]

One further non-Gaussian example is given by

$$C(h)=\exp[-a\int |\tilde{h}(k)|^2dk-b\int \tilde{W}(k)|\tilde{h}(k)|^\alpha dk],$$

where $1\leq\alpha<2$. Now

$$\mathcal{V}_C=\{h:\int |\tilde{h}(k)|^2dk<\infty,\int \tilde{W}(k)|\tilde{h}(k)|^\alpha dk<\infty\},$$

which is a normed space but not a Hilbert space. Suppose next that $\tilde{W}(k)\to|k|^{-\sigma}$, $0<1-\frac{1}{3}\sigma<\frac{1}{2}\alpha<1$, and choose T such that

$$\tilde{h}_T(k)=S^d\tilde{h}(Sk),$$

where $d=(3-\sigma)/\alpha<\frac{3}{2}$. Then

$$C(T^nh)\to C^{(T)}(h)=\exp[-b\int |k|^{-\sigma}|\tilde{h}(k)|^\alpha dk]$$

and it follows that

$$\mathcal{V}_C{}^{(T)}=\{h:\int |k|^{-\sigma}|\tilde{h}(k)|^\alpha dk<\infty\}.$$

From the metric convergence point of view $d_C(h)$ is equivalent to the norm $\|h\|$, where here

$$\|h\|^2\equiv\int |\tilde{h}(k)|^2dk+[\int \tilde{W}(k)|\tilde{h}(k)|^\alpha dk]^{2/\alpha}.$$

The sequence of norms $\|T^nh\|$ converges to yield

$$\|T^nh\|^2\to\|h\|^2_{(T)}=[\int |k|^{-\sigma}|\tilde{h}(k)|^\alpha dk]^{2/\alpha},$$

which is invariant under T. It is interesting to add that when $0<\alpha<1$, $d_C(h)$ is not equivalent to *any* norm, but nevertheless the general argument still holds.

As another observation on this example observe that if $\alpha=2(1-\frac{1}{3}\sigma)$, $\sigma>0$, then $d=\frac{3}{2}$, and one learns that

$$C^{(T)}(h)=\exp[-a\int |\tilde{h}(k)|^2dk-b\int |k|^{-\sigma}|\tilde{h}(k)|^\alpha dk];$$

namely, both terms survive in this "canonical" scaling limit. Clearly, however, $\mathcal{V}_C{}^{(T)}\neq L^2(R^3)$, and the distinction can be easily exhibited by invoking a second transformation, say T_1, where $\tilde{h}_{T_1}(k)\equiv S^{3/2}\tilde{h}(S(k-\kappa))$, $\kappa\neq 0$. In this case one finds that

$$C^{(T_1)}(h)=\exp[-a\int |\tilde{h}(k)|^2dk],$$

independently of the variable b.

When $\alpha<2$, the normal concept of correlation length loses its meaning for essentially any choice of $\tilde{W}(k)$. Long-range correlation persists, say, even if $\tilde{W}(k)=1$, as can be seen from a lack of clustering of $C(h)$. Choose $h(x)=h_1(x)+h_2(x+a)$ and study the limit as $a\to\infty$ of $C(h)$. Since this expression does not ultimately tend to $C(h_1)C(h_2)$, one simply cannot ask for the details of such convergence as needed to determine the correlation length.

To exhibit yet another feature more clearly, let $\tilde{W}(k)=1$ and consider

$$C(h)=\exp[-a\int |\tilde{h}(k)|^2dk-b\int |\tilde{h}(k)|^\alpha dk],$$

for which

$$\mathcal{V}_C=\{h:\int [|\tilde{h}(k)|^2+|\tilde{h}(k)|^\alpha]dk<\infty\}.$$

Under a scaling limit where $\tilde{h}_T(k)=S^{3/\alpha}\tilde{h}(Sk)$ with $0<S<1$ (ultraviolet *not* infrared) one finds that

$$C^{(T)}(h)=\exp[-b\int |\tilde{h}(k)|^\alpha dk]$$

and

$$\mathcal{V}_C{}^{(T)}=\{h:\int |\tilde{h}(k)|^\alpha dk<\infty\},$$

which, as generally is the case, contains *new* elements not in $\mathcal{V}_C$. One such element is clearly

$$\tilde{h}(k)=|k|^{-\beta}\exp(-k^2),\quad \alpha\beta<3\leq 2\beta,$$

which if α is small means β can be relatively large determining thereby an unusual but valid element $h(x)$ (a generalized function) possessing long-range tails.

Physically, it is interesting to compare the present situation to that when σ is discrete. In the latter case, the physical system responds to test variables $[C(\lambda\bar{h})\neq 1$ for some $\lambda]$ to which it gave *no* response previously $[C(\lambda\bar{h})=1$ for all $\lambda]$, and in this way a revised ordering could be determined. In the present case, the physical system *saturates* $[C(\lambda\bar{h})=0$ for all $\lambda\neq 0]$ for some test variables and any potential revised ordering that may exist is completely overwhelmed by some other uninteresting interaction. Probing first in the opposite extreme (ultraviolet) leads to a system that no longer exhibits the same saturation levels (as if the offending interaction has been switched off). Afterwards one finds the secondary behavior revealed and certain very long-range test variables are allowed that were prohibited previously.

In this note we have attempted to relate some general principles of the renormalization group by means of various transformations on the characteristic function and to emphasize the general change arising in the space of allowed test variables when a phase boundary is reached or crossed. Although our examples have been largely Gaussian, the concepts, of course, are generally applicable. Moreover, it should be em-

phasized that the random variable Φ could include a multitude of *dependent* variables, which the measure μ would properly relate, and which would make fairly general correlations as easy to consider in principle as the simplest ones. Specifically, for the spin sequence example, we could imagine that

$$\Phi \equiv (\{\Phi_{k_1}\}, \{\Phi_{k_1}\Phi_{k_2}\}, \{\Phi_{k_1}\Phi_{k_2}\Phi_{k_3}\}, \cdots),$$
$$h \equiv (\{h_{k_1}\}, \{h_{k_1k_2}\}, \{h_{k_1k_2k_3}\}, \cdots),$$

and that μ ensured that the variable $\Phi_{k_1}\Phi_{k_2}$ was always the product of the two variables Φ_{k_1} and Φ_{k_2}, etc., simply by appropriate δ functions. Any pair correlation is adequately described by the quadratic terms in $\ln C(h)$, and according to the above description, such quadratic terms could in principle contain any desired correlation of interest.

Furthermore, in close analogy to the linearized analysis near the critical point used for Hamiltonians,[1] one can imagine linearization of the effects of (infinitesimal) renormalization group transformations *directly on* $\ln C(h)$ in the neighborhood of a fixed point $\ln C^{(T)}(h)$. By comparing the expansion coefficients of $\ln[C_T(h)/C^{(T)}(h)]$ and $\ln[C(h)/C^{(T)}(h)]$, when developed in Volterra series, one can determine, for an infinitesimal transformation T, a linear operator L_C that acts on the Volterra coefficients in the manner of a generator, at least in principle. This operator, too, has relevant and irrelevant eigenvalues,[1] the former leading away from critical behavior, the latter representing universality. The discussion presented in this note is, however, not aimed at developing calculational techniques for specific problems, but rather at simply presenting a unifying overview of some basic renormalization group ideas in the general framework of characteristic functions.

[1]See, e.g., K.G. Wilson and J.B. Kogut, Phys. Rep. C **12**, 75 (1974).

[2]G. Jona-Lasinio, Nuovo Cimento B **26**, 99 (1975).

[3]Although not explicitly used, one simple bound on the support of μ often applies. Assume that positive numbers α_k and test variables $h_{(k)}$, $k=1,2,\ldots$, exist such that (i)$\Sigma\, \alpha_k (h_{(k)}, h)^2 \geq 0$, equality holding only if $h \equiv 0$, and (ii)$\Sigma\, \alpha_k \int (h_{(k)}, \Phi)^2 d\mu(\Phi) < \infty$. Then μ is supported on the space of random variables $\{\Phi : \Sigma\, \alpha_k (h_{(k)}, \Phi)^2 < \infty\}$.

[4]G. C. Hegerfeldt and J.R. Klauder, Commun. Math. Phys. **16**, 329 (1970). The arguments given therein are readily specialized to cover the present case. For simplicity, we assume that nonzero nonfunctional test variables for which $C(\lambda h) \equiv 1$ for all real λ are excluded.

[5]The idealized expression $C(h) = \exp(-a\Sigma\, h_k^2 + 2ib\Sigma\, h_k)$ is probably more familiar below the critical temperature than the one in the text. But for these two choices of characteristic function the metrics are equivalent, and the spaces of allowed test variables are identical.

[6]E. Brezin, J.C. LeGuillon, and J. Zinn-Justin, "Field Theoretical Approach to Critical Phenomena," in *Phase Transitions and Critical Phenomena, Vol.* 6, edited by C. Domb and M.S. Green (Academic, New York, to be published).

[7]A. Aharony, "Dependence of Universal Critical Behavior on Symmetry and Range of Interaction," in Ref. 6.

[8]J.R. Klauder, Acta Physica Austr. Suppl. VIII, 227 (1971); in *Mathematical Methods in Theoretical Physics*, Lectures in Theoretical Physics, Vol. XIVB, edited by W.E. Britten (Colorado Associated U.P., Boulder, 1974), p. 329.

On the nonuniqueness of the Lagrangian

Seán Browne

Dublin Institute for Advanced Studies, Dublin 4, Ireland
(Received 19 August 1975)

We describe a general method for comparing two theories which are based on Lagrangians which differ only by the divergence of a 4-vector term, and we examine how the generators of symmetry transformations depend on this 4-divergence term. For a Poincaré invariant theory we find that the Poincaré generators are independent of this term. Corresponding results are obtained for the case of internal symmetries and for scale and conformal symmetry.

I. INTRODUCTION

It is well known that the Lagrangian of a system of fields is not unique since we may add to it a 4-divergence term without altering the equations of motion. Also any continuous invariance of the action integral yields by way of Noether's theorem[1] a divergenceless current, allowing us to construct a time-independent quantity which may be identified as the generator of the transformation in question. The question thus arises as to whether, for a fixed given invariance of the action integral, changing the Lagrangian in this way alters the corresponding generator.[2] We shall see in fact that it does. There now immediately arises the disquieting prospect that the S Matrix depends on this 4-divergence term. But we shall see in fact that it does not.

The aim of the present paper is to examine in some detail the exact dependence of the currents and generators on this 4-divergence term. We shall consider Lagrangians which contain in general derivatives of the basic fields higher than just the first derivatives. Besides the generality that this affords is the important fact[3] that the correct Lagrangian for the usual free massless scalar field *must* contain the second derivatives of the field if we are to have an action integral which is invariant under special conformal transformations. We warn the reader at this point that invariance here means strict invariance of the action integral and not just invariance up to a 4-divergence, although we shall comment upon this latter case later on. The occurrence of higher order derivatives in the Lagrangian turns out in fact to bear crucially upon the conformal properties of the system of fields.

Our main result is easily stated: In a Poincaré invariant theory the Poincaré generators are independent of the 4-divergence term whereas the scale and special conformal generators are not, in general. Thus, in particular, the S matrix is independent of this term. We also examine the exact circumstances under which the scale and conformal generators are independent of this 4-divergence term.

The tone of this paper is completely classical and our results, in the form presented here, are true only for classical fields and, possibly, free quantum fields. We particularly wish to emphasize this remark with respect to our treatment of scale and conformal symmetry. The trouble here arises from the fact that, even though a given classical Lagrangian is scale invariant and contains no masses or dimensional coupling constants, the corresponding quantized theory does contain a dimensional parameter. This parameter must be introduced to perform the subtractions necessary to renormalize the theory and render it finite.[4] We shall also assume throughout that the scale dimensions of the fields are just their canonical dimensions as determined by dimensional analysis. We do this despite the fact that such appears highly unlikely in any interesting theory as can be very easily seen using the renormalization group equations.[5]

The problems associated with quantizing higher order Lagrangian theories have been discussed by many authors,[6] and more recently by Riewe and Green,[7] where many additional references may be found. However, there seems to be no completely satisfying solution to those problems at the present time. For free Lagrangians of the aforementioned type we may also use the quantization technique of Takahashi and Umezawa.[8] An interesting point about this method is that it works directly from the field equations and is thus manifestly invariant under the addition of a 4-divergence to the corresponding Lagrangian. We shall not discuss these questions further here.

The organization of the paper is as follows: Sec. 2 reviews briefly higher order Lagrangians. In Sec. 3 we develop a method for comparing theories which are based on Lagrangians which are related through the addition of a 4-divergence term. We apply this method in Secs. 4,5, and 6, to determine how the Poincaré, scale, and special conformal generators, respectively, depend on the Lagrangian used to construct them. Section 7 contains a summary and conclusions.

We use the Pauli metric with $x_4 = ict$ and units where $\hbar = c = 1$.

2. NONUNIQUENESS OF CURRENTS AND LAGRANGIANS

Consider a system of fields $\phi_a^{(\kappa)}(x)$ $(\kappa = 1,2, \ldots, l)$ described by a Lagrangian $\mathcal{L}$, which depends on x only through the fields $\phi_a^{(\kappa)}(x)$ and their first, say, n derivatives. Variation of the action of the system $W_{21} = \int_{\sigma_1}^{\sigma_2} \mathcal{L}(x)\, d^4x$ yields the equation of motion in the usual way. Invariance of W_{21} under the infinitesimal transformation

$$x_\mu \to x'_\mu = x_\mu + \delta x_\mu,$$
$$\phi_a^{(\kappa)}(\chi) \to \phi_a^{(\kappa)\prime}(x'), \tag{2.1}$$

where $\phi_a^{(\kappa)\prime}(x)=\phi_a^{(\kappa)}(x)+\delta\phi_a^{(\kappa)}(x)$, provides us with the divergenceless current[3]

$$J_\mu^{(c)}=\sum_{\kappa=1}^{l}\sum_{r=0}^{n-1}\Pi^{(\kappa)}_{a,\mu(r)\mu}\delta\phi^{(\kappa)}_{a,\mu(r)}+\delta x_\mu \mathcal{L}, \tag{2.2}$$

where

$$\Pi^{(\kappa)}_{a,\mu(r)}=\sum_{\kappa=1}^{l}\sum_{s=0}^{n-r}(-1)^s\partial_{\lambda(s)}\left(\frac{\partial\mathcal{L}}{\partial\phi^{(\kappa)}_{a,\mu(r)\lambda(s)}}\right).$$

$J_\mu^{(c)}(x)$ calculated in this manner is called the canonical current associated with the transformation (2.1) and $G(\sigma)=\int_\sigma J_\mu^{(c)}\,d\sigma_\mu$ can be identified with the generator of the transformation (2.1) for the fields $\phi_a^{(\kappa)}(x)$. Furthermore, if we add to $J_\mu^{(c)}$ the divergence of an antisymmetric tensor, then neither the divergence of the current nor its corresponding generator are altered.

Besides this nonuniqueness of the current there is another well-known nonuniqueness, namely, that of the Lagrangian itself. If we replace $\mathcal{L}$ by another Lagrangian $\tilde{\mathcal{L}}=\mathcal{L}+\partial_\mu\Lambda_\mu$, where $\Lambda_\mu(x)$ depends on x only through the fields and their derivatives, then the equations of motion derived from $\tilde{\mathcal{L}}$ are the same as those derived from $\mathcal{L}$. However, the canonical current (2.2) associated with $\tilde{\mathcal{L}}(x)$ is in general different from the corresponding current associated with $\mathcal{L}(x)$ leading possibly to different generators for the same transformation—a disquieting prospect. In the following sections we shall investigate the relationship between these two currents and between their corresponding generators.

3. AN ASSOCIATED EQUIVALENT CURRENT

The two types of nonuniqueness discussed in the previous section are of an essentially independent nature. The nonuniqueness of the current presents no problems and is well understood. In fact it is precisely this nonuniqueness that allows us to construct the many[9] symmetric energy momentum tensors from a given Lagrangian. The corresponding nonuniqueness of the Lagrangian is not so well understood, and we may ask, for example, whether or not the energy–momentum tensors are insensitive to this nonuniqueness. Should, however, the two currents $J_\mu^{(c)}$ and $\tilde{J}_\mu^{(c)}$ differ only by the divergence of an antisymmetric tensor, then the question of the nonuniqueness of the Lagrangian becomes either redundant or trivial. Unfortunately, this is not the case.

For the general infinitesimal transformation (2.1) we have, using the usual technique of adding and subtracting the same quantity, that

$$\Delta\tilde{\mathcal{L}}(x)+\partial_\mu(\delta x_\mu)\tilde{\mathcal{L}}(x)=\partial_\mu\left(J_\mu^{(c)}(x)+\delta\Lambda_\mu(x)+\delta x_\mu\partial_\alpha\Lambda_\alpha(x)\right), \tag{3.1}$$

where $\Delta\mathcal{L}(x)\equiv\mathcal{L}'(x')-\mathcal{L}(x)$ and $\mathcal{L}'(x)\equiv\mathcal{L}(\phi'(x))$. From (3.1) it follows[3,10] that if the action integral $\tilde{W}_{21}=\int_{\sigma_1}^{\sigma_2}\tilde{\mathcal{L}}(x)\,d^4x$ is invariant under the infinitesimal transformation (2.1), then the current

$$\tilde{J}_\mu=\tilde{J}_\mu^{(c)}+\delta\Lambda_\mu+\delta x_\mu\partial_\alpha\Lambda_\alpha \tag{3.2}$$

is divergenceless. Notice that we have written $\tilde{J}_\mu$ rather than $\tilde{J}_\mu^{(c)}$. The reason for this is that in constructing $\tilde{J}_\mu$ we did not use the "canonical procedure" described in the previous section—instead we used the fact $\partial_\mu(\delta\Lambda_\mu)=\delta(\partial_\mu\Lambda_\mu)$. We have thus no reason to suspect that $\tilde{J}_\mu$ is the same as the canonical current $\tilde{J}_\mu^{(c)}$ calculated from $\tilde{\mathcal{L}}(x)$ using Eq. (2.2). In fact these two currents $\tilde{J}_\mu$ and $\tilde{J}_\mu^{(c)}$ are *not* the same. However, they differ only by the divergence of an antisymmetric tensor. If this were not the case, then $\tilde{J}_\mu$ and $\tilde{J}_\mu^{(c)}$ could lead to different generators for the same transformation. This would drastically limit the usefulness of $\tilde{J}_\mu$ in examining the relationship between theories based on $\mathcal{L}$ and $\tilde{\mathcal{L}}$, despite its simple relationship with $\tilde{J}_\mu^{(c)}$, if we wished to remain within the so-called "canonical framework".

The relationship between $\tilde{J}_\mu$ and $\tilde{J}_\mu^{(c)}$ is given by

$$\tilde{J}_\mu=\tilde{J}_\mu^{(c)}+\partial_\nu n_{\mu\nu}, \tag{3.3}$$

where

$$n_{\mu\nu}=\sum_{\kappa=1}^{l}\sum_{s=0}^{m-1}\sum_{r=0}^{m-s-1}\frac{(-1)^r(s+1)}{(r+s+2)}\partial_{\mu(r)}\left\{\frac{\partial\Lambda_\mu}{\partial\phi^{(\kappa)}_{a,\mu(r)\lambda(s)\nu}}-\frac{\partial\Lambda_\nu}{\partial\phi^{(\kappa)}_{a,\mu(r)\lambda(s)\mu}}\right\}\delta\phi^{(\kappa)}_{a,\lambda(s)} \tag{3.4}$$

as can be verified by explicitly evaluating $\partial_\nu n_{\mu\nu}$. The "m" which appears in the summation in (3.4) is the highest field derivative in $\partial_\mu\Lambda_\mu$. The condition that the two currents $\tilde{J}_\mu$ and $\tilde{J}_\mu^{(c)}$ are equal can be read off Eq. (3.4); it is

$$\frac{\partial\Lambda_\alpha}{\partial\phi^{(\kappa)}_{a,\mu(r)\beta}}=\frac{\partial\Lambda_\beta}{\partial\phi^{(\kappa)}_{a,\mu(r)\alpha}} \tag{3.5}$$

for all $r\geq 0$. There are two particularly interesting forms of Λ_μ for which Eq. (3.5) holds automatically. These are:

(A) Λ_μ depends only on the fields $\phi_a^{(\kappa)}$ and not their derivatives.

(B) $\Lambda_\mu=\partial_\lambda\sigma_{\mu\lambda}$, where $\sigma_{\mu\nu}=\sigma_{\nu\mu}$ and $\sigma_{\mu\nu}$ depends only on the fields $\phi_a^{(\kappa)}$ and not their derivatives.

The proof of these statements is straightforward. Case (A) occurs when we restrict our attention to Lagrangians which depend only on the fields and their first derivatives. This is the most usual case. Case(B) is more or less peculiar to Lagrangians which are "nearly conformally invariant".[11] These are Lagrangians which can be made conformally covariant by adding a suitable 4-divergence term.

The generator associated with the current $\tilde{J}_\mu^{(c)}$, or equivalently $\tilde{J}_\mu$, is

$$\begin{aligned}\tilde{G}(\sigma)&=\int_\sigma\tilde{J}_\mu(x)\,d\sigma_\mu(x)\\&=G(\sigma)+\int_\sigma(\delta\Lambda_\mu)+\delta x_\mu\partial_\alpha\Lambda_\alpha)\,d\sigma_\mu\end{aligned} \tag{3.6}$$

where $G(\sigma)$ is the generator associated with the current $J_\mu^{(c)}$, which is derived from the Lagrangian, $\mathcal{L}(x)$. It follows from Eq. (3.6) that, when a term of the form $\partial_\mu\Lambda_\mu$ is added to the Lagrangian, the generator corresponding to the arbitrary infinitesimal transformation (2.1) picks up the additional contribution

$$\Delta G(\sigma)=\int_\sigma(\delta\Lambda_\mu+\delta x_\mu\partial_\alpha\Lambda_\alpha)\,d\sigma_\mu. \tag{3.7}$$

It is immediately obvious from this equation that in the case of an internal symmetry, i.e., when $\delta x_\mu=0$ in the transformation (2.1), that $G(\sigma)$ picks up the definite contribution

$$\Delta G(\sigma)=\int_\sigma \delta\Lambda_\mu\, d\sigma_\mu \tag{3.8}$$

when we alter the Lagrangian by adding to it a general term of the form $\partial_\mu\Lambda_\mu$. Thus the generator of internal symmetry transformations is independent of the (invariant) Lagrangian used to construct it if $\Delta G(\sigma)=0$ for every Λ_μ which satisfies $\delta(\partial_\mu\Lambda_\mu)=0$. For a general Lagrangian we do not know if the condition $\delta(\partial_\mu\Lambda_\mu)=0$ is sufficient to guarantee $\Delta G(\sigma)=0$. For first order Lagrangians, however, we shall now prove that the condition $\Delta G(\sigma)=0$ does indeed hold.

Since the Lagrangian is restricted to depend only on the fields and their first derivatives, Λ_μ can depend only on the fields. Thus the condition $\delta(\partial_\mu\Lambda_\mu)=\partial_\mu(\delta\Lambda_\mu)=0$ is simply

$$\frac{\partial}{\partial\phi_i}(\delta\Lambda_\mu)\phi_{i,\mu}=0,$$

where, for convenience, we have put all the fields into the single vector ϕ. Differentiating this equation with respect to $\phi_{i,\mu}$ we obtain

$$\frac{\partial}{\partial\phi_i}(\delta\Lambda_\mu)=0$$

so that $\delta\Lambda_\mu$ is just a constant, independent of the fields. The assumed Lorentz transformation properties of Λ_μ now gives

$$\delta\Lambda_\mu=0$$

so that $\Delta G(\sigma)=0$, as already stated.

In the remaining sections we shall examine the currents $\tilde J_\mu$ and $\tilde J_\mu^{(c)}$ and the quantity $\Delta G(\sigma)$ for the case where $\delta x_\mu\neq 0$ in (2.2). We shall further limit the δx_μ to general conformal transformations only, which consists of Poincaré, scale, and special conformal transformations respectively. The 15-parameter conformal group is the largest group of transformations in Minkowski space which preserves the metric relation $ds^2=0$.

4. DEPENDANCE OF THE POINCARE GENERATORS ON THE LAGRANGIAN

Under the infinitesimal Poincaré transformation defined by

$$\begin{aligned}&\delta x_\mu=\epsilon_\mu+\omega_{\mu\nu}x_\nu,\quad \omega_{\mu\nu}=-\omega_{\nu\mu},\\ &\phi_a^{(\kappa)\prime}(x')=\phi_a^{(\kappa)}(x)+\tfrac{1}{2}i\omega_{\mu\nu}(S_{\mu\nu}^{(\kappa)})_{ab}\phi_b^{(\kappa)},\end{aligned} \tag{4.1}$$

where $S_{\mu\nu}^{(\kappa)}$ is the spin tensor of the field $\phi_a^{(\kappa)}(x)$, we shall assume that the Lagrangian transforms according to $\mathcal{L}(x)\to\mathcal{L}'(\chi')=\mathcal{L}(\chi)$ which guarantees the invariance of the action integral W_{21}, under the transformation (4.1). We can now construct the corresponding conserved quantities using equation (2.3). These are the canonical energy momentum tensor

$$T_{\mu\nu}^{(c)}=\mathcal{L}\delta_{\mu\nu}-\sum_{\kappa=1}^{l}\sum_{r=0}^{n-1}\Pi_{a,\mu(r)\mu}^{(\kappa)}\phi_{a,\mu(r)\nu}^{(\kappa)} \tag{4.2}$$

and the canonical momentum tensor

$$\mathcal{M}_{\mu\sigma\lambda}^{(c)}=x_\sigma T_{\mu\lambda}^{(c)}-x_\lambda T_{\mu\sigma}^{(c)}-2F_{\mu\sigma\lambda}. \tag{4.3}$$

The explicit form of the quantity $F_{\mu\sigma\lambda}$, which appears in Eqs. (4.3), is irrelevant for our purposes and we omit it. Corresponding to $T_{\mu\nu}^{(c)}$ and $\mathcal{M}_{\mu\sigma\lambda}^{(c)}$ we have the Poincaré generators $P_\mu=\int_\sigma T_{\lambda\mu}^{(c)}\,d\sigma_\lambda$ and $M_{\sigma\lambda}=\int_\sigma \mathcal{M}_{\mu\sigma\lambda}^{(c)}\,d\sigma_\mu$.

Let us now turn our attention to the equivalent Poincaré invariant Lagrangian $\tilde{\mathcal{L}}$. In accordance with our discussion in the previous section, there are quantities $\tilde T_{\mu\nu}$ and $\tilde{\mathcal{M}}_{\mu\sigma\lambda}^{(c)}$, which differ from $\tilde T_{\mu\nu}^{(c)}$ and $\tilde{\mathcal{M}}_{\mu\sigma\lambda}^{(c)}$, respectively, only by the divergences of two antisymmetric tensors, and which are related to $T_{\mu\nu}^{(c)}$ and $M_{\mu\sigma\lambda}^{(c)}$ through Eq. (3.2). After a little algebra we find that $T_{\mu\nu}$ and $T_{\mu\nu}^{(c)}$ differ only by the divergence of an antisymmetric tensor. Thus $\tilde P_\mu=P_\mu$. In a similar way we find that $\tilde M_{\sigma\lambda}=M_{\sigma\lambda}$.

We have thus shown that the Poincaré generators are independent of the (Poincaré invariant, of course) Lagrangian used to construct them. In view of this result we may take the Poincaré currents to be

$$\theta_{\mu\nu}\quad\text{and}\quad x_\sigma\theta_{\mu\lambda}-x_\lambda\theta_{\mu\sigma}, \tag{4.4}$$

where $\theta_{\mu\nu}$ is a symmetric energy–momentum tensor constructed from $\mathcal{L}(x)$ or from any equivalent Lagrangian $\mathcal{L}(x)$ of the form $\tilde{\mathcal{L}}=\mathcal{L}+\partial_\mu\Lambda_\mu$. It follows from this statement that the Hamiltonian, and hence, the S matrix, is also independent of the term $\partial_\mu\Lambda_\mu$.

5. DEPENDANCE OF THE SCALE GENERATOR OF THE LAGRANGIAN

In general the Lagrangian will depend on masses $m^{(\kappa)}$, corresponding to the fields $\phi_a^{(\kappa)}$ $(\kappa=1,2,\ldots,l)$, and dimensional coupling constants f_i $(i=1,2,\ldots,q)$. The scale transformation we shall consider here is constructed in such a way that scale invariance of the action integral corresponds to the vanishing of the masses and dimensional coupling constants from the Lagrangian. Infinitesimally, this is given by[12]

$$\delta x_\mu=\epsilon x_\mu,\quad \phi_a^{(\kappa)\prime}(x')=\phi_a^{(\kappa)}(x)+\epsilon l^{(\kappa)}\phi_a^{(\kappa)}(x), \tag{5.1}$$

where $l^{(\kappa)}$ is the length dimension of $\phi^{(\kappa)}$ in units where $\hbar=c=1$ and the corresponding current is

$$S_\mu^{(c)}=xT_{\mu\nu}^{(c)}+G_\mu \tag{5.2}$$

where

$$G_\mu=\sum_{\kappa=1}^{l}\sum_{r=0}^{n-1}(l^{(\kappa)}-r)\Pi_{a,\mu(r)\mu}^{(\kappa)}\phi_{a,\mu(r)}^{(\kappa)}. \tag{5.3}$$

The current $S_\mu^{(c)}$ is divergenceless whenever all masses and dimensional coupling constants are absent from the Lagrangian. In fact we have

$$\partial_\mu S_\mu^{(c)}=\Delta(m,f,\mathcal{L}), \tag{5.4}$$

where, if η_i is the length dimension of f_i $(i=1,2,\ldots,q)$,

$$\Delta(m,f,\mathcal{L})=\sum_{\kappa=1}^{l}m^{(\kappa)}\frac{\partial\mathcal{L}}{\partial m^{(\kappa)}}-\sum_{i=1}^{q}\eta_i f_i\frac{\partial\mathcal{L}}{\partial f_i}. \tag{5.5}$$

Turning our attention now to the equivalent Lagrangian $\tilde{\mathcal{L}}$, we obtain

$$\begin{aligned}\tilde S_\mu=S_\mu^{(c)}+\sum_{\kappa=1}^{l}\sum_{r=0}^{m}\frac{\partial\Lambda_\mu}{\partial\phi_{a,\mu(r)}^{(\kappa)}}(l^{(\kappa)}-r-x_\alpha\partial_\alpha)\phi_{a,\mu(r)}^{(\kappa)}\\ +x_\mu\partial_\alpha\Lambda_\alpha.\end{aligned} \tag{5.6}$$

Also we have, by dimensional analysis,

$$-3\Lambda_\mu=\sum_{\kappa=1}^{I}\sum_{\tau=0}^{m}\frac{\partial\Lambda_\mu}{\partial\phi_{a,\mu(\tau)}^{(\kappa)}}(l^{(\kappa)}-\tau)\phi_{a,\mu(\tau)}^{(\kappa)}-\Delta(m,f,\Lambda_\mu),\quad(5.7)$$

where $\Delta(m,f,\Lambda_\mu)$ is given by Eq. (5.5) but with $\mathcal{L}$ replaced by Λ_μ. Combining Eqs. (5.6) and (5.7), we obtain

$$\tilde{S}_\mu=S_\mu^{(c)}+\partial_\alpha(x_\mu\Lambda_\alpha-x_\alpha\Lambda_\mu)+\Delta(m,f,\Lambda_\mu).\quad(5.8)$$

The generator corresponding to $\tilde{S}_\mu$ is

$$\tilde{D}(\sigma)=\int_\sigma\tilde{S}_\mu\,d\sigma_\mu$$

$$=D(\sigma)+\int\Delta(m,f,\Lambda_\mu)\,d\sigma_\mu,\quad(5.9)$$

where $D(\sigma)$ is the generator corresponding to $S_\mu^{(c)}$. The second term on the right-hand side of Eq. (5.9) brings out clearly the explicit dependence of the scale generator $\tilde{D}(\sigma)$ on the masses and dimensional coupling constants occurring in Λ_μ. Also, as we would expect, $D(\sigma)$ and $\tilde{D}(\sigma)$ are different in general for a Poincaré invariant theory. In the limit of scale invariance, however, the last term in (5.10) vanishes and the scale generator is indeed independent of the Lagrangian used to construct it.

As a further example of the results of Sec. 3 we shall consider, briefly, the possibility of writing the scale current as a moment of the canonical energy–momentum tensor for some Lagrangian. For simplicity we shall not consider this question in its complete generality. Instead we shall assume that Λ_μ satisfies Eq. (3.5). This, as we have already noted, guarantees that $\tilde{J}_\mu=\tilde{J}_\mu^{(c)}$, without too much loss in generality. We begin by combining (4.2) and (5.8) to obtain

$$\tilde{S}_\mu^{(c)}=x_\nu\tilde{T}_{\mu\nu}^{(c)}+G_\mu-3\Lambda_\mu+\Delta(m,f,\Lambda_\mu),\quad(5.10)$$

where G_μ is given by Eq. (5.3). Equation (5.10) says that if we choose Λ_μ in such a way that

$$G_\mu-3\Lambda_\mu+\Delta(m,f,\Lambda_\mu)=0,$$

then we may write $\tilde{S}_\mu^{(c)}=x_\nu\tilde{T}_{\mu\nu}^{(c)}$ as required. Thus, for example, if G_μ does not depend on any masses or other dimensional coupling constants, then by changing to the equivalent Lagrangian

$$\tilde{\mathcal{L}}=\mathcal{L}+\tfrac{1}{3}\partial_\mu G_\mu\quad(5.11)$$

we can write $\tilde{S}_\mu^{(c)}=x_\nu\tilde{T}_{\mu\nu}^{(c)}$. Whether or not G_μ depends on any dimensional quantities, we may always write $\tilde{S}_\mu^{(c)}=x_\nu\tilde{T}_{\mu\nu}^{(c)}$ in the limit of scale invariance by changing to the Lagrangian (5.11).

Let us apply this to the free massless scalar field ϕ described by $\mathcal{L}=-\frac{1}{2}(\partial_\mu\phi)(\partial_\mu\phi)$. Here $G_\mu=\phi\partial_\mu\phi$ so that by changing to

$$\tilde{\mathcal{L}}=-\tfrac{1}{2}(\partial_\mu\phi)(\partial_\mu\phi)+\tfrac{1}{3}\partial_\mu(\phi\partial_\mu\phi)\quad(5.12)$$

we may write $\tilde{S}_\mu^{(c)}=x_\nu\tilde{T}_{\mu\nu}^{(c)}$, where $\tilde{T}_{\mu\nu}^{(c)}=-\frac{1}{2}\partial_\alpha\phi\partial_\alpha\phi\delta_{\mu\nu}+\partial_\mu\phi\partial_\nu\phi-\frac{1}{6}(\partial_\mu\partial_\nu-\delta_{\mu\nu}\Box)\phi^2$. Considerations of the present kind involving changing the Lagrangian in order to reorganize the scale current in terms of the canonical energy–momentum tensor have been given by Macfarlane[13] and Takahashi.[14]

Finally we comment that the scalar field Lagrangian (5.12), besides serving as an example in this highly contrived situation, has absolutely no physical significance.

6. DEPENDENCE OF THE SPECIAL CONFORMAL GENERATORS ON THE LAGRANGIAN

The infinitesimal special conformal transformation is defined by the relations[12]

$$\delta x_\mu=c_\lambda a_{\mu\lambda},\quad a_{\mu\lambda}=2x_\mu x_\lambda-\delta_{\mu\lambda}x^2,$$
$$\phi_a^{(\kappa)\prime}(x')=\phi_a^{(\kappa)}(x)+2c_\lambda(l^{(\kappa)}x_\lambda-ix_\sigma S_{\lambda\sigma}^{(\kappa)})_{ab}\phi_b^{(\kappa)}(x),\quad(6.1)$$

where $S_{\mu\nu}^{(\kappa)}$ and $l^{(\kappa)}$ are the spin tensor and length dimension respectively, of the field $\phi_a^{(\kappa)}$. The current here is the canonical special conformal tensor

$$K_{\mu\lambda}^{(c)}=a_{\lambda\rho}T_{\mu\rho}^{(c)}+2x_\lambda G_\mu+4x_\sigma F_{\mu\sigma\lambda}+W_{\mu\lambda}.\quad(6.2)$$

Of the quantities G_μ, $F_{\mu\sigma\lambda}$, and $W_{\mu\lambda}$ occurring in (6.2), G_μ and $F_{\mu\sigma\lambda}$ have already been mentioned, while the explicit form of $W_{\mu\lambda}$ as a function of the fields and their derivatives is unimportant for our purposes.

From Eq. (3.2) we see that $\tilde{J}_\mu$ contains a term $\delta\Lambda_\mu$, which is the infinitesimal change induced in this case by the transformation (6.1), and which may be written as[11]

$$\delta\Lambda_\mu=c_\lambda\{2(-3x_\lambda-ix_\sigma S_{\lambda\sigma})_{\mu\alpha}\Lambda_\alpha-a_{\sigma\lambda}\partial_\sigma\Lambda_\mu$$
$$+2x_\lambda\Delta(m,f,\Lambda_\mu)+V_\lambda(\Lambda_\mu)\},\quad(6.3)$$

where $S_{\mu\nu}$ is the spin tensor for a vector field and $V_\lambda(\Lambda_\mu)$ is a function of the fields and their derivatives, and whose explicit form is again not needed here. What *is* important about $V_\lambda(\Lambda_\mu)$ is that, in a scale invariant theory, Λ_μ transforms like a vector field under special conformal transformations if and only if $V_\lambda(\Lambda_\mu)$ vanishes. At any rate we find, using (3.2) and (6.3), that

$$\tilde{K}_{\mu\lambda}=K^{(c)}+\partial_\sigma(a_{\mu\lambda}\Lambda_\sigma-a_{\sigma\lambda}\Lambda_\mu)+2x_\lambda\Delta(m,f,\Lambda_\mu)+V_\lambda(\Lambda_\mu).\quad(6.4)$$

Thus the generators corresponding to $\tilde{K}_{\mu\lambda}$ are

$$\tilde{K}_\lambda(\sigma)=K_\lambda(\sigma)+2\int_\sigma x_\lambda\Delta(m,f,\Lambda_\mu)\,d\sigma_\mu+\int_\sigma V_\lambda(V_\mu)\,d\sigma_\mu,\quad(6.5)$$

where $K_\lambda(\sigma)$ are the special conformal generators corresponding to $K_{\mu\lambda}^{(c)}$. It follows easily from Eq. (6.5) that if we add a general term of the form $\partial_\mu\Lambda_\mu$ to a Lagrangian, $\mathcal{L}$, whose action integral is invariant under special conformal transformations, then the resulting Lagrangian's action, $\tilde{W}_{21}$, will in general not be invariant under special conformal transformations. We have already come across this point in our discussion of the spin-0 field.[3] It also follows from Eq. (6.5) that even if both of the action integrals W_{21} and $\tilde{W}_{21}$ (i.e., corresponding to $\mathcal{L}$ and $\tilde{\mathcal{L}}$ respectively) are invariant under special conformal transformations, the special conformal generators K_λ and $\tilde{K}_\lambda$ will in general be different unless the last term on the right-hand side of Eq. (6.5) vanishes.[5] In particular this term will vanish if Λ_μ transforms covariantly under special conformal transformations so that $V_\lambda(\Lambda_\mu)=0$. This is a stronger condition than the invariance condition $V_\lambda(\partial_\mu\Lambda_\mu)=0$ and we do not know if it holds for all Λ_μ. What we do know is that this condition $[V_\lambda(\Lambda_\mu)=0]$ holds for a large class of Lagrangians including all first order Lagrangians. This follows easily from the explicit form of $V_\lambda(\Lambda_\mu)$ which is given in Ref. 11. For this class of Lagrangians it may also

be shown that if any one of these Lagrangians yields an action integral invariant under special conformal transformations then *all* do. Furthermore, all these Lagrangians give rise to the same special conformal generators as can be easily seen from (6.5).

7. SUMMARY AND CONCLUSIONS

We have presented a method for comparing two theories which are based on Lagrangians differing only by a 4-divergence term. For internal symmetries we found, at least for first order Lagrangians, that the generators did not depend on the particular Lagrangian we used to construct them, while the Poincaré generators were *in general* independent of the (Poincaré invariant) Lagrangian used to construct them. This latter result has the consequence that the Poincaré currents could be taken as moments of a symmetric energy–momentum tensor constructed from any of these equivalent Lagrangians.

For scale and special conformal transformations we found simple expressions for the change in the generators due to the addition of the 4-divergence term to the Lagrangian. It then followed that in the limit of scale invariance the scale generator was independent of the particular Lagrangian used to construct it. Although for a large class of Lagrangians [those related by a Λ_μ, for which $V_\lambda(\Lambda_\mu)$ is the divergence of an antisymmetric tensor] the special conformal generators were independent of the Lagrangian in the limit of full conformal symmetry, we were not able to prove this result for arbitrary Λ_μ. For Lagrangians which depend on the fields and their first derivatives only, our results state that for invariance under internal symmetry transformations the generators are independent of the particular Lagrangian used to construct them and that in the limit of *scale* invariance all 15 conformal generators are independent of the Lagrangian used to construct them.

Finally, let us remark that most of our results may be transferred in an obvious way to the case where the notion of invariance corresponds to the Lagrangian being invariant only up to the addition of an arbitrary 4-divergence term, i.e., invariance of the integrated Lagrangian. Perhaps this notion of invariance will turn out to be the more important one in the end.[16]

[1]E. Noether, Goett. Nachr., 235 (1918).
[2]J. Schwinger, Phys. Rev. 82, 914 (1951).
[3]S. Browne, Proc. Roy. Ir. Acad. A 73, 179 (1973).
[4]S. Coleman and R. Jackiw, Ann. Phys. (N.Y.) 67, 552 (1971).
[5]G. Attarelli, Nuovo Cimento 4, 335 (1974). See p. 365.
[6]T.S. Chang, Proc. Roy. Cambridge Phil. Soc. 44, 76 (1948); J.S. deWet, Proc. Roy. Cambridge Phil. Soc. 44, 546 (1948); S.P. Misra, Indian J. Phys. 33, 461, 520 (1959); A.O. Barut and G.H. Mullen, Ann. Phys. (N.Y.) 20, 203 (1962).
[7]F. Riewe and A.E.S. Green, J. Math. Phys. 13, 1368 (1972). See also H.P. Dürr, Nuovo Cimento A 22, 386 (1974).
[8]See, for example, Y. Takahashi, *Introduction to Field Quantization* (Pergamon, New York, 1969), p. 112.
[9]F.J. Bellinfante, Physica 7, 449 (1940); E. Huggins, Ph.D. Thesis, Cal. Tech. (1962); C.G. Callan Jr., S. Coleman, and J. Jackiw, Ann. Phys. (N.Y.) 59, 42 (1970).
[10]E.M. Corson, *Introduction to Tensors, Spinors and Relativistic Wave Equations* (Hafner, New York, 1953).
[11]S. Browne, Proc. Roy. Ir. Acad. A 74, 49 (1974).
[12]J. Wess, Nuovo Cimento 18, 1086 (1960); G. Mack and A. Salam, Ann. Phys. (N.Y.) 53, 174 (1969).
[13]A.J. Macfarlane, Univ. of Cambridge preprint, 1970.
[14]Y. Takahashi, Phys. Rev. D 3, 622 (1971).
[15]In general, if $\delta\tilde{W}_{21}=\delta W_{21}$ for all σ_1,σ_2, then, by (3.1), $\delta(\partial_\alpha\Lambda_\alpha)+\partial_\mu(\delta x_\mu\partial_\alpha\Lambda_\alpha)=0$. This implies that $\Delta G(\sigma)$ is independent of σ. It does not, however, imply that $\Delta G(\sigma)=0$. See P. Roman, Nuovo Cimento 10, 546 (1958).
[16]J. Wess and B. Zumino, Nucl. Phys. B 70, 39 (1974); A. Salam and J. Strathdee, Phys. Rev. D 11, 1521 (1975).

Bäcklund transformations and the equation $z_{xy} = F(x,y,z)$

S. G. Byrnes

Department of Mathematics, University of Durham, England
(Received 11 August 1975)

It will be shown that the only equations of the form $z_{xy} = F(x,y,z)$ which possess Bäcklund transformations to take one solution of this equation into another solution of the same equation are either linear or else can be obtained from the sin(h)–Gordon equation (or $z_{xy} = e^z$) by a simple change of scale and/or a displacement of the dependent variable.

1. INTRODUCTION

Certain nonlinear equations are known to possess solutions which are generally referred to as solitons.[1] What one means by this is roughly as follows: If one thinks of the independent variables as labeling position then a soliton is a localized disturbance which has the property of "retaining its shape" on interaction with another soliton. It may occupy a position different to what it would have without the interaction. Of the equations which are known to possess solitons only one is real and of second order. As most of physics seems to be reasonably well described by second-order equations one is tempted to ask if this equation—the sin(*h*)-Goron equation—is the only real second-order equation to possess solitons.

McLaughlin and Scott[2] have answered this question in relation to equations of the form $z_{xy} = F(z)$. This paper extends their result to equations of the form $z_{xy} = F(x, y, z)$. The main problem in this area is how to extend the soliton concept to higher dimensions, i.e., to include more independent variables. Is the soliton a two-dimensional object or does its counterpart exist in higher dimensions? Hirota[3] has obtained solutions of the sine-Gordon equation in (2 + 1)-dimensions which he calls solitons. However these are "plane wave" solutions, and are not localized at a point. To extend the soliton concept to higher than (1 + 1)-dimensions it seems desirable to look for localized solutions. Even so, the question of what corresponds to Bäcklund transformations in higher dimensions remains to be answered.

Recently the theory of solitons has created a great deal of interest within elementary particle theory. The field satisfying the sine-Gordon equation in (1 + 1) dimensions, i.e.,

$$\frac{\partial^2\phi}{\partial x^2} - \frac{\partial^2\phi}{\partial t^2} + \frac{\mu^2}{\beta}\sin(\beta\phi) = 0,$$

where μ^2 and β are constants, t is the time and x the position, has been successfully quantized.[4] The result is that in addition to the "usual particles" obtained in such theories there are particles which correspond to "quantized solitons." The usual particles are obtained from the free field equation $\phi_{xx} - \phi_{tt} + \mu^2\phi = 0$ by perturbation, treating the "interaction potential" $-(1/4!)\mu^2\beta^2\phi^4 + (1/6!)\mu^2\beta^4\phi^6 + \cdots$ as small. These particles satisfy Bose statistics whereas the quantized soliton satisfy Fermi statistics. So one has found that solitons do, in this theory, correspond to particles when quantized. They give a richer spectrum (i.e., more particles) than one would expect from an analogy with the simple harmonic oscillator.

Throughout this paper the name "sin(*h*)-Gordon equation" will be used to refer to all real equations of the form $\phi_{xy} = F(\phi)$ where F is a function of a single variable and satisfies $F''(z) = KF(z)$ for some constant $K \neq 0$. So for real constants A, k and ϵ, $F(z)$ must have one of the following forms: $A\sin(kz+\epsilon)$, $A\sinh(kz+\epsilon)$, $A\cosh(kz+\epsilon)$, or $A\exp(kz)$. Here as elsewhere in this paper primes denote derivatives with respect to the variable displayed; hence $F''(z)$ means the second derivative of the function $F(z)$ with respect to z. Also the x and y subscripts denote derivatives with respect to the variables, so $\phi_{xy} = \partial^2\phi/\partial x\partial y$.

The concept of a soliton given above is rather vague and one would like a more rigorous definition. The concept of a Bäcklund transformation seems to provide this. Recently it has been shown by Lamb[5] that every equation which is known to have soliton solutions also possesses a Bäcklund transformation which takes every solution of the given equation into another solution of the same equation. So it seems reasonable to define an equation to have the soliton property if it possesses a Bäcklund transformation of this type.[6]

An equation of the form

$$\frac{\partial^2 z}{\partial x\partial y} = F\left(x, y, z, \frac{\partial z}{\partial x}, \frac{\partial z}{\partial y}, \frac{\partial^2 z}{\partial x^2}, \frac{\partial^2 z}{\partial y^2}\right), \tag{1.1}$$

where the function F is analytic in each of its variables, will be said to possess a Bäcklund transformation if there exist functions $P(x,y,z,w,p,q)$ and $Q(x,y,z,w,p,q)$ which are analytic in each variable and which satisfy the following condition for *all* solutions $z(x,y)$ of the Eq. (1.1).

If $w(x,y)$ is any solution of the coupled set of equations

$$\begin{aligned} w_x &= P(x,y,z,w,z_x,z_y), \\ w_y &= Q(x,y,z,w,z_x,z_y), \end{aligned} \tag{1.2}$$

for the given solution $z(x,y)$ of (1.1), then $w(x,y)$ satisfies the original equation, i.e., $w_{xy} = F(x,y,w,w_x,w_y, w_{xx}, w_{yy})$.

Note that (1.2) assumes a particular form for the Bäcklund transformation. In principle the functions P and Q could depend on higher derivatives of z and w and also on integrals of them. For example, P and Q in (1.2) could depend on $z_{xx}, z_{yyy}, \int z_x \cdot dy$ as well as the variables considered there. An objection to the use of integrals in the Bäcklund transformation is the observa-

tion that the equation $z_{xy}=F(x,y,z)$, for any function F, possesses the "Bäcklund transformation," i.e.,

$$w_x=\int[F(x,y,w)-\alpha\cdot F(x,y,z)]\circ dy+\alpha\circ z_x,$$
$$w_y=\int[F(x,y,w)-\beta\cdot F(x,y,z)]\circ dx+\beta\circ z_y,$$

where α and β are arbitrary nonzero constants.

Now consider the case of F linear in the dependent variables, i.e., in (1.1) take

$$F=A_1+A_2\circ z+A_3\cdot z_x+A_4\circ z_y+A_5\circ z_{xx}+A_6\cdot z_{yy},$$

where the A_i $(i=1,\ldots,6)$ may be functions of x and y but not of z or its derivatives. With this F, Eq. (1.1) trivially has a Bäcklund transformation of the type (1.2) which may be found as follows. If $\alpha(x,y)$ is any solution of the homogeneous equation

$$z_{xy}=A_2\cdot z+A_3\cdot z_x+A_4\cdot z_y+A_5\cdot z_{xx}+A_6\cdot z_{yy},$$

then for *all* solutions $z(x,y)$ of (1.1) and for all constants K, one has that $w(x,y)=z(x,y)+K\circ\alpha(x,y)$ is also a solution of (1.1), since F is linear. Dividing by $\alpha(x,y)$ and differentiation gives the Bäcklund transformation

$$\frac{\partial w}{\partial x}=\frac{\partial z}{\partial x}+\frac{1}{\alpha}\cdot\frac{\partial\alpha}{\partial x}\cdot(w-z),$$
$$\frac{\partial w}{\partial y}=\frac{\partial z}{\partial y}+\frac{1}{\alpha}\cdot\frac{\partial\alpha}{\partial y}\cdot(w-z). \tag{1.3}$$

For the sin(h)-Gordon equation

$$z_{xy}=A\exp(kz)+B\exp(-kz), \tag{1.4}$$

where $k\neq0$, A and B are (possibly complex) constants, the Bäcklund transformation is

$$\frac{\partial w}{\partial x}=\frac{\partial z}{\partial x}+(2a/k)(A\exp[k(w+z)/2]+B\exp[-k(w+z)/2]), \tag{1.5}$$

$$\frac{\partial w}{\partial y}=-\frac{\partial z}{\partial y}+(1/a)(\exp[k(w+z)/2]-\exp[-k(w-z)/2]), \tag{1.6}$$

where a is a constant. Note that (1.5) and (1.6) are nontrivial in the sense that one cannot set $\phi=w-z$ and obtain two equations for ϕ independent of z and w; whereas the linear case (1.3) is trivial.

One sees immediately from (1.4), (1.5), and (1.6) that if one changes scale, i.e., replaces x everywhere by some function of x, i.e., $x=x(x')$ say, and y by some function of y, i.e., $y=y(y')$ say (and uses the new variables x' and y') then one obtains another equation which possesses a Bäcklund transformation. More generally one may replace x and y by new independent variables u and v where $x=x(u,v)$ and $y=y(u,v)$ for arbitrary functions; however the equation which then possesses the Bäcklund transformations will not be of the form $z_{xy}=F(x,y,z)$. Further, one sees from (1.4)–(1.6) that if one displaces the dependent variable, i.e., if one replaces z and w everywhere by $z+\alpha$ and $w+\alpha$ respectively for some function $\alpha(x,y)$ then one obtains another equation which possesses a Bäcklund transformation. More generally if $f(x,y,\phi)$ is any function of three variables then one may replace z and w everywhere by $f(x,y,z)$ and $f(x,y,w)$ respectively to obtain other equations which possess Bäcklund transformations.

It seems worth recording that there are nonlinear equations which possess trivial Bäcklund transformations as discussed above. Let A_0, A_1, A_3, A_4, A_5, A_6, and K be functions of x and y and let g be any function of y only. Suppose that A_i $(i=0,1,3,\ldots,6)$ and K satisfy the following conditions:

$$A_4(x,y)\circ\{g''(y)+[g'(y)]^2\}+A_1(x,y)+A_3(x,y)g'(y)=0, \tag{1.7}$$

$$\frac{\partial^2K}{\partial x^2}\neq0, \tag{1.8}$$

$$A_1\circ K+g'(y)\frac{\partial K}{\partial x}+A_3\circ\frac{\partial K}{\partial y}+A_4\circ\frac{\partial^2K}{\partial x^2}+A_5\circ\frac{\partial^2K}{\partial y^2}=\frac{\partial^2K}{\partial x\partial y}. \tag{1.9}$$

Then the equation

$$\frac{\partial^2z}{\partial x\partial y}=A_0+A_{1z}+g'(y)\frac{\partial z}{\partial x}+A_3\frac{\partial z}{\partial y}+A_4\frac{\partial^2z}{\partial x^2}+A_5\frac{\partial^2z}{\partial y^2}+G\left(\frac{\partial^2z}{\partial x^2}\right) \tag{1.10}$$

possess a Bäcklund transformation if $G(r)$ is any function of period $\partial^2K/\partial x^2$, e.g., $G(r)=\sin(2\pi r/K_{xx})$. Let $f(z)$ be an arbitrary function of a single variable and define a function $\alpha(z,x)$ of two variables, by the equation

$$z=x\circ\alpha(z,x)+f[\alpha(z,x)]. \tag{1.11}$$

The Bäcklund transformation for (1.10) is then

$$\frac{\partial w}{\partial x}=\exp[g(y)]\alpha[\exp(-g(y))(w-z-K),x]+\frac{\partial K}{\partial x}+\frac{\partial z}{\partial x}, \tag{1.12}$$

$$\frac{\partial w}{\partial y}=g'(y)(w-z-K)+\frac{\partial K}{\partial y}+\frac{\partial z}{\partial y}. \tag{1.13}$$

Note that for a given $z(x,y)$ the general solution of (1.12) and (1.13) is

$$w=z+K+[ax+f(a)]\exp[g(y)],$$

where a is an arbitrary constant.

The rest of this paper is concerned with showing that if the equation $z_{xy}=F(x,y,z)$ possesses a Bäcklund transformation then it is the sin(h)-Gordon equation up to a change of scale and a displacement of the dependent variable.[7] For a physically interesting theory it is reasonable to impose Lorentz invariance. In the coordinates chosen this means that the equation $\phi_{xy}=F$ should be invariant under the replacement of x and y by λx and $(1/\lambda)y$ respectively for some constant $\lambda\neq0$. However, the imposition of Lorentz invariance does not seem to simplify the problem and so will not be considered further.

2. THE BASIC EQUATIONS

Consider the equation

$$\frac{\partial^2z}{\partial x\partial y}=F(x,y,z). \tag{2.1}$$

Suppose this equation possesses a Bäcklund transformation of the type discussed in the introduction. Differen-

tiating the first equation in (1.2) with respect to y and the second with respect to x and demanding that both z and w satisfy (2.1) gives

$$F(x,y,w)=\frac{\partial P}{\partial y}+\frac{\partial P}{\partial z}z+\frac{\partial P}{\partial w}Q+\frac{\partial P}{\partial p}F(x,y,z)+\frac{\partial P}{\partial q}t, \quad (2.2)$$

$$F(x,y,w)=\frac{\partial Q}{\partial x}+\frac{\partial Q}{\partial z}p+\frac{\partial Q}{\partial w}P+\frac{\partial Q}{\partial p}r+\frac{\partial Q}{\partial p}F(x,y,z), \quad (2.3)$$

where $p=z_x$, $q=z_y$, $r=z_{xx}$, $s=z_{xy}$, and $t=z_{yy}$. Since $z(x,y)$ is any solution of (2.1) one may treat x,y,z,p,q, r, and t in (2.2) and (2.3) as independent variables. Further since $w(x,y)$ is any solution of (2.2) and (2.3) one may take w in (2.2) and (2.3) as an extra independent variable. One now treats (2.2) and (2.3) as two partial differential equations for $P(x,y,z,w,p,q)$ and $Q(x,y,z,w,p,q)$ where x,y,z,w,p,q,r, and t are all independent variables.

The case of $F(x,y,z)$ linear in z has already been treated in the introduction so in all of what follows it will be assumed that $F(x,y,z)$ is not linear in z. In the following the linear case arises when one derives an equation of the form

$$\frac{\partial F}{\partial z}=k_1\cdot\frac{\partial F}{\partial w}+k_2, \quad (2.4)$$

where k_1 and k_2 are functions of x and y (but not z or w) and $\partial F/\partial z=(\partial/\partial z)\{F(x,y,z)\}$. Now the right hand side of (2.4) is independent of z and the left hand side is independent of w so both must equal some function, k_3 say, of x and y only. Integrating $\partial F/\partial z=k_3$ then gives $F(x,y,z)=k_3(x,y)z+k_4(x,y)$ for some function k_4 of x and y. But this F is then linear in the "dependent" variable. This argument will not be repeated.

Now differentiate (2.2) w.r.t. t and (2.3) w.r.t. r,

$$\frac{\partial P}{\partial q}=0 \quad \text{and} \quad \frac{\partial Q}{\partial p}=0,$$
$$\therefore P=P(x,y,z,w,p), \quad (2.5)$$
$$Q=Q(x,y,z,w,q).$$

Differentiate (2.2) twice w.r.t. q and use the fact that P is independent of q to obtain

$$\frac{\partial P}{\partial w}\cdot\frac{\partial^2 Q}{\partial q^2}=0. \quad (2.6)$$

If $\partial P/\partial w=0$ then one may differentiate (2.2) with respect to w to obtain $\partial F/\partial w=0$, which is a contradiction since F is assumed to be nonlinear. Hence (2.6) implies $\partial^2 Q/\partial q^2=0$.

$$\therefore Q=Q_0(x,y,z,w)+Q_1(x,y,z,w)q, \quad (2.7)$$

on integrating where Q_0 and Q_1 are functions of x,y,z, and w which are to be determined. Similarly on differentiating (2.3) twice w.r.t. p, one obtains for some functions P_0 and P_1 that

$$P=P_0(x,y,z,w)+P_1(x,y,z,w)p. \quad (2.8)$$

Substituting (2.7) and (2.8) into (2.2) and (2.3) and equating coefficients of pq, p, q and terms independent of p and q gives the following eight equations:

$$\frac{\partial P_1}{\partial w}Q_1+\frac{\partial P_1}{\partial z}=0, \quad (2.9)$$

$$\frac{\partial P_1}{\partial w}Q_0+\frac{\partial P_1}{\partial y}=0, \quad (2.10)$$

$$\frac{\partial P_0}{\partial w}Q_1+\frac{\partial P_0}{\partial z}=0, \quad (2.11)$$

$$\frac{\partial P_0}{\partial w}Q_0+\frac{\partial P_0}{\partial y}+P_1\circ F(x,y,z)=F(x,y,w), \quad (2.12)$$

$$\frac{\partial Q_1}{\partial w}P_1+\frac{\partial Q_1}{\partial z}=0, \quad (2.13)$$

$$\frac{\partial Q_0}{\partial w}P_1+\frac{\partial Q_0}{\partial z}=0, \quad (2.14)$$

$$\frac{\partial Q_1}{\partial w}P_0+\frac{\partial Q_1}{\partial x}=0, \quad (2.15)$$

$$\frac{\partial Q_0}{\partial w}P_0+\frac{\partial Q_0}{\partial x}+Q_1\cdot F(x,y,z)=F(x,y,w)\ . \quad (2.16)$$

Note that throughout this paper, equality of functions will be used in the sense of "identically equal to." So, for example $P_1\neq 0$ means that $P_1(x,y,z,w)$ is not zero everywhere, although it may be zero at a point, or points.

The rest of this section is devoted to proving that

$$P_1\neq 0 \quad \text{and} \quad Q_1\neq 0. \quad (2.17)$$

Now one must demand that the Bäcklund transformation must transform from one solution to another, i.e., that it must "depend on z." So the transformation from z to w,

$$\frac{\partial w}{\partial x}=P(x,y,w), \quad \frac{\partial w}{\partial y}=Q(x,y,w), \quad (2.18)$$

for some functions P and Q is not allowed since it is independent of z and its derivatives (i.e., p and q).

The proof of (2.17) proceeds by contradiction so suppose that

$$P_1=0. \quad (2.19)$$

Equations (2.13) and (2.14) then give

$$\frac{\partial Q_1}{\partial z}=0=\frac{\partial Q_0}{\partial z}. \quad (2.20)$$

Differentiate (2.15) and (2.16) w.r.t. z and use (2.20)

$$\frac{\partial Q_1}{\partial w}\cdot\frac{\partial P_0}{\partial z}=0, \quad (2.21)$$

$$\frac{\partial Q_0}{\partial w}\circ\frac{\partial P_0}{\partial z}+Q_1\circ\frac{\partial F}{\partial z}=0. \quad (2.22)$$

Now if $\partial P_0/\partial z=0$ then (2.22) gives $Q_1=0$ since F is not linear. But then the Bäcklund transformation is of the form (2.18). Hence (2.21) implies

$$\frac{\partial Q_1}{\partial w}=0. \quad (2.23)$$

Equations (2.15), (2.20), and (2.23) then give

$$Q_1=Q_1(y). \quad (2.24)$$

Take $Q_1\partial/\partial w+\partial/\partial z$ of (2.12) and use (2.11) and (2.20):

$$\left(Q_1\frac{\partial Q_0}{\partial w}-Q_1'(y)\right)\frac{\partial P_0}{\partial w}=Q_1\cdot\frac{\partial F}{\partial w}. \tag{2.25}$$

Use (2.11) to replace $\partial P_0/\partial z$ by $\partial P_0/\partial w$ in (2.22),

$$Q_1\cdot\frac{\partial Q_0}{\partial w}\cdot\frac{\partial P_0}{\partial w}=Q_1\cdot\frac{\partial F}{\partial z}. \tag{2.26}$$

Subtract (2.26) from (2.25),

$$-Q_1'(y)\frac{\partial P_0}{\partial w}=Q_1\frac{\partial F}{\partial w}-Q_1\frac{\partial F}{\partial z}. \tag{2.27}$$

Take $Q_1\partial/\partial w+\partial/\partial z$ of (2.27) and use (2.11), i.e.,

$$Q_1^2\frac{\partial^2 F}{\partial w^2}=Q_1\frac{\partial^2 F}{\partial z^2}. \tag{2.28}$$

Note that if both $P_1=0$ and $Q_1=0$ then (2.11) and (2.14) show that both P_0 and Q_0 are independent of z. But then the Bäcklund transformation is of the form (2.18) which is not allowed. Hence $Q_1\neq 0$ and (2.28) gives

$$Q_1\frac{\partial^2 F}{\partial w^2}=\frac{\partial^2 F}{\partial z^2}=A(x,y), \tag{2.29}$$

since the lhs of (2.28) is independent of z and the rhs is independent of w.

Now $A\neq 0$ since F is not linear. Hence (2.29) implies $Q_1=1$. But then (2.27) implies that $\partial F/\partial w=\partial F/\partial z$ which is a contradiction since F is not linear.

Hence the result (2.17) is established.

3. THAT P_1 AND Q_1 ARE CONSTANTS AND $P_1+Q_1=0$

Suppose that

$$\frac{\partial P_1}{\partial w}\neq 0. \tag{3.1}$$

Apply the operator $Q_1\partial/\partial w+\partial/\partial z$ to (2.10):

$$\left(Q_1\frac{\partial Q_0}{\partial w}+\frac{\partial Q_0}{\partial z}-Q_0\frac{\partial Q_1}{\partial w}-\frac{\partial Q_1}{\partial y}\right)\frac{\partial P_1}{\partial y}+\left(Q_0\frac{\partial}{\partial w}+\frac{\partial}{\partial y}\right)\left(Q_1\frac{\partial P_1}{\partial w}+\frac{\partial P_1}{\partial z}\right)=0.$$

But because of (2.9) the second term in this equation is zero and from (3.1) one may write

$$Q_1\frac{\partial Q_0}{\partial w}+\frac{\partial Q_0}{\partial z}-Q_0\frac{\partial Q_1}{\partial w}-\frac{\partial Q_1}{\partial y}=0. \tag{3.2}$$

Now take $Q_1\partial/\partial w+\partial/\partial z$ of (2.12),

$$\frac{\partial P_0}{\partial w}\left(Q_1\frac{\partial Q_0}{\partial w}+\frac{\partial Q_0}{\partial z}-Q_0\frac{\partial Q_1}{\partial w}-\frac{\partial Q_1}{\partial y}\right)+\left(Q_0\frac{\partial}{\partial w}+\frac{\partial}{\partial y}\right)\left(Q_1\frac{\partial P_0}{\partial w}+\frac{\partial P_0}{\partial z}\right)+P_1\frac{\partial F}{\partial z}=Q_1\frac{\partial F}{\partial w}. \tag{3.3}$$

The first term here is zero by (3.2) and the second by (2.11). Hence

$$P_1\frac{\partial F}{\partial z}=Q_1\frac{\partial F}{\partial w}. \tag{3.4}$$

Now if also

$$\frac{\partial Q_1}{\partial w}\neq 0, \tag{3.5}$$

then in a similar way one may apply the operator $P_1\partial/\partial w+\partial/\partial z$ to (2.15) and (2.16) obtaining

$$Q_1\frac{\partial F}{\partial z}=P_1\frac{\partial F}{\partial w}. \tag{3.6}$$

Now from Eqs. (3.4) and (3.6)

$$Q_1\left(\frac{\partial F}{\partial z}\right)^2=\left(Q_1\frac{\partial F}{\partial z}\right)\frac{\partial F}{\partial z}=\left(P_1\frac{\partial F}{\partial w}\right)\frac{\partial F}{\partial z}$$

$$=\frac{\partial F}{\partial w}\left(P_1\frac{\partial F}{\partial z}\right)=\frac{\partial F}{\partial w}\left(Q_1\frac{\partial F}{\partial w}\right)=Q_1\left(\frac{\partial F}{\partial w}\right)^2. \tag{3.7}$$

But from (2.17), $Q_1\neq 0$, so (3.7) implies that

$$\left(\frac{\partial F}{\partial z}\right)^2=\left(\frac{\partial F}{\partial w}\right)^2. \tag{3.8}$$

But (3.8) gives $F_z=\pm F_w$, i.e., F is linear which is a contradiction.

Now the basic Eqs. (2.9) to (2.16) are symmetric under the substitution $P_i\leftrightarrow Q_i$ $(i=0,1)$ and $x\leftrightarrow y$. So it is sufficient to suppose that (3.1) is true but (3.5) is not. That is, suppose

$$\frac{\partial P_1}{\partial w}\neq 0 \quad\text{and}\quad \frac{\partial Q_1}{\partial w}=0. \tag{3.9}$$

Equations (3.9), (2.13) and (2.15) imply

$$Q_1=Q_1(y). \tag{3.10}$$

Take $Q_1\partial/\partial w+\partial/\partial z$ of (2.16) and use (2.11), (3.2) and (3.10), i.e.,

$$Q_1\partial F/\partial z=Q_1\partial F/\partial w. \tag{3.11}$$

But $Q_1\neq 0$ by (2.17) so (3.11) implies that F is linear which is a contradiction. Hence (3.9) cannot hold. Similarly the case $\partial P_1/\partial w=0$, $\partial Q_1/\partial w\neq 0$ cannot hold. So it has been shown that

$$\frac{\partial P_1}{\partial w}=0 \quad\text{and}\quad \frac{\partial Q_1}{\partial w}=0. \tag{3.12}$$

Equations (3.12), (2.9), (2.10), (2.13) and (2.15) imply

$$P_1=P_1(x) \quad\text{and}\quad Q_1=Q_1(y). \tag{3.13}$$

The rest of this section is concerned with proving that $P_1+Q_1=0$. The proof is by contradiction, so suppose that

$$P_1+Q_1\neq 0. \tag{3.14}$$

Note that expressions like $P_1(x)+Q_1(y)=0$ or $P_1(x)\cdot Q_1(y)=1$ imply that both P_1 and Q_1 are constants since x and y are independent variables.

Use (2.11) and (3.13) in (3.3) and eliminate $\partial P_0/\partial z$ in favor of $\partial P_0/\partial w$.

$$(Q_1-P_1)\frac{\partial P_0}{\partial w}\frac{\partial Q_0}{\partial w}-Q_1'(y)\frac{\partial P_0}{\partial w}=Q_1\frac{\partial F}{\partial w}-P_1\frac{\partial F}{\partial z}. \tag{3.15}$$

Similarly taking $P_1\partial/\partial w+\partial/\partial z$ of (2.16) and using (2.14),

$$(P_1-Q_1)\frac{\partial P_0}{\partial w}\frac{\partial Q_0}{\partial w}-P_1'(x)\frac{\partial Q_0}{\partial w}=P_1\frac{\partial F}{\partial w}-Q_1\frac{\partial F}{\partial z}. \tag{3.16}$$

Add (3.15) to (3.16), i.e.,

$$P_1'(x)\frac{\partial Q_0}{\partial w}+Q_1'(y)\frac{\partial P_0}{\partial w}+(P_1+Q_1)\frac{\partial F}{\partial w}-\frac{\partial F}{\partial w}=0. \tag{3.17}$$

Now if P_1 and Q_1 are both constants then (3.17) implies that $P_1+Q_1=0$ since F is not linear, i.e., $F_w \neq F_z$. So from (3.17) one deduces

$$[P_1'(x)=0 \text{ and } Q_1'(y)=0] \text{ imply } P_1+Q_1=0. \tag{3.18}$$

Take $(P_1\partial/\partial w+\partial/\partial x)(Q_1\partial/\partial w+\partial/\partial z)$ of (3.15) and (3.16) and use (2.11) and (2.14),

$$(P_1-Q_1)^3\frac{\partial^2 P_0}{\partial w^2}\frac{\partial^2 Q_0}{\partial w}=Q_1^2P_1\frac{\partial^3 F}{\partial w^3}-P_1\frac{\partial^3 F}{\partial z^3} \tag{3.19}$$

$$-(P_1-Q_1)^3\frac{\partial^2 P_0}{\partial w^2}\frac{\partial^2 Q_0}{\partial w^2}=P_1^2Q_1\frac{\partial^3 F}{\partial w^3}-Q_1\frac{\partial^3 F}{\partial z^3}. \tag{3.20}$$

Add (3.19) to (3.20) and use (3.14),

$$P_1Q_1\frac{\partial^3 F}{\partial w}=\frac{\partial^3 F}{\partial z^3}. \tag{3.21}$$

Now (3.21) implies that F_{zzz} is independent of z, so $F_{zzz}=F_{www}$. If $F_{zzz}\neq 0$ then (3.21) implies that $P_1(x)Q_1(y)=1$, which implies that both P_1 and Q_1 are constants; (3.18) then gives $P_1+Q_1=0$ in contradiction to (3.14), so

$$\frac{\partial^3 F}{\partial z^3}=0. \tag{3.22}$$

Then (3.19) implies

$$\frac{\partial^2 P_0}{\partial w^2}\frac{\partial^2 Q_0}{\partial w^2}=0, \tag{3.23}$$

since $P_1(x)-Q_1(y)=0$ again, it says that P_1 and Q_1 are constants which is, via (3.18), a contradiction to (3.14). Because of the symmetry between the P's and Q's one may, from (3.23) take

$$\frac{\partial^2 P_0}{\partial w^2}=0. \tag{3.24}$$

Take $P_1\partial/\partial w+\partial/\partial z$ of (3.16) and use (2.14) and (3.24),

$$P_1^2\frac{\partial^2 F}{\partial w^2}-Q_1\frac{\partial^2 F}{\partial z^2}=0. \tag{3.25}$$

Now (3.25) implies that F_{zz} is independent of z. If $F_{zz}=0$ then F is linear which is a contradiction. Hence $F_{ww}=F_{zz}\neq 0$. So (3.25) implies that $[P_1(x)]^2=Q_1(y)$. But then P_1 and Q_1 are constants in contradiction to (3.14) upon using (3.18).

Hence the result is established, i.e., P_1 and Q_1 are constants and

$$P_1+Q_1=0. \tag{3.26}$$

4. THAT $F_{zzz}=K\cdot F_z$ AND $P_1^2=1$

From the previous section one may write

$$P_1=-Q_1=c, \tag{4.1}$$

where c is a constant. From (2.17), $c\neq 0$. Take $-c\partial/\partial w+\partial/\partial z$ of (2.12),

$$2\frac{\partial P_0}{\partial w}\frac{\partial Q_0}{\partial w}=\frac{\partial F}{\partial z}+\frac{\partial F}{\partial w}. \tag{4.2}$$

Take $c\partial/\partial w+\partial/\partial z$ of (4.2) and use (2.11) and (2.14),

$$4c\frac{\partial^2 P_0}{\partial w^2}\frac{\partial Q_0}{\partial w}=c\frac{\partial^2 F}{\partial w^2}+\frac{\partial^2 F}{\partial z^2}. \tag{4.3}$$

Take $c\,\partial/\partial w+\partial/\partial z$ of (2.12) and use (2.11) and (2.14), i.e.,

$$2c\frac{\partial^2 P_0}{\partial w^2}Q_0+2c\frac{\partial^2 P_0}{\partial w\partial y}+c\frac{\partial F}{\partial z}=c\frac{\partial F}{\partial w}. \tag{4.4}$$

If $\partial^2 P_0/\partial w^2=0$ then take $\partial/\partial w$ of (4.4) to obtain $c\partial^2 F/\partial w^2=0$ which is a contradiction since F is not linear.

$$\therefore\quad \frac{\partial^2 P_0}{\partial w^2}\neq 0. \tag{4.5}$$

Similarly

$$\frac{\partial^2 Q_0}{\partial w^2}\neq 0. \tag{4.6}$$

Eliminate $\partial Q_0/\partial w$ from (4.2) and (4.3),

$$2c\frac{\partial^2 P_0}{\partial w^2}\left(\frac{\partial F}{\partial z}+\frac{\partial F}{\partial w}\right)=\frac{\partial P_0}{\partial w}\left[c\frac{\partial^2 F}{\partial w^2}+\frac{\partial^2 F}{\partial z^2}\right]. \tag{4.7}$$

Take $-c\partial/\partial w+\partial/\partial z$ of (4.7) and use (2.11), i.e.,

$$2c\frac{\partial^2 P_0}{\partial w^2}\left[-c\frac{\partial^2 F}{\partial w^2}+\frac{\partial^2 F}{\partial z^2}\right]=\frac{\partial P_0}{\partial w}\left[-c^2\frac{\partial^3 F}{\partial w^3}+\frac{\partial^3 F}{\partial z^3}\right]. \tag{4.8}$$

Take $-c\partial/\partial w+\partial/\partial z$ of (4.8) and use (2.11),

$$2c\frac{\partial^2 P_0}{\partial w^2}\left(c^2\frac{\partial^3 F}{\partial w^3}+\frac{\partial^3 F}{\partial z^3}\right)=\frac{\partial P_0}{\partial w}\left(c^3\frac{\partial^4 F}{\partial w^4}+\frac{\partial^4 F}{\partial z^4}\right). \tag{4.9}$$

Take $-c\partial/\partial w+\partial/\partial z$ of (4.2) and use (2.11) and (2.14),

$$-4c\frac{\partial P_0}{\partial w}\frac{\partial^2 Q_0}{\partial w^2}=-c\frac{\partial^2 F}{\partial w^2}+\frac{\partial^2 F}{\partial z^2}. \tag{4.10}$$

Now if $\partial^4 F/\partial z^4=0$ then (4.9) implies $c^2\partial^3 F/\partial w^3=-\partial^3 F/\partial z^3$ since by (4.5), $\partial^2 P_0/\partial w^2\neq 0$. If $c^2\neq 1$, then the equation $c^2\partial^3 F/\partial w^3=-\partial^3 F/\partial z^3$ must imply that $\partial^3 F/\partial z^3=0$. But then (4.8) implies that $-c(\partial^2 F/\partial w^2)+(\partial^2 F/\partial z^2)=0$ since $\partial^2 P_0/\partial w^2\neq 0$. Then (4.10) implies $-4c(\partial P_0/\partial w)/(\partial^2 Q/\partial w)=0$ which contradicts (4.5), (4.6), or $c\neq 0$.

$$\therefore\quad \frac{\partial^4 F}{\partial z^4}\neq 0 \text{ implies } \quad c^2=1. \tag{4.11}$$

Now from Eqs. (4.7) and (4.8)

$$2c\frac{\partial^2 P_0}{\partial w^2}\frac{\partial P_0}{\partial w}\left(\frac{\partial F}{\partial z}+\frac{\partial F}{\partial w}\right)\left(-c^2\frac{\partial^3 F}{\partial w^3}+\frac{\partial^3 F}{\partial z^3}\right)$$
$$=2c\frac{\partial^2 P_0}{\partial w^2}\frac{\partial P_0}{\partial w}\left(c\frac{\partial^2 F}{\partial w^2}+\frac{\partial^2 F}{\partial z^2}\right)\left(-c\frac{\partial^2 F}{\partial w^2}+\frac{\partial^2 F}{\partial z^2}\right).$$

But from (4.5), $\partial^2 P_0/\partial w\neq 0$. So $\partial P_0/\partial w\neq 0$, and

$$\left(c^2\frac{\partial^3 F}{\partial w^3}-\frac{\partial^3 F}{\partial z^3}\right)\left(\frac{\partial F}{\partial z}+\frac{\partial F}{\partial w}\right)=\left(c\frac{\partial^2 F}{\partial w^2}\right)^2-\left(\frac{\partial^2 F}{\partial z^2}\right)^2. \tag{4.12}$$

Take $\partial^2/\partial w\partial z$ of (4.12),

$$c^2\frac{\partial^4 F}{\partial w^4}\frac{\partial^2 F}{\partial z^2}=\frac{\partial^2 F}{\partial w^2}\frac{\partial^4 F}{\partial z^4}. \tag{4.13}$$

Because of (4.11) one then has

$$\frac{\partial^4 F}{\partial z^4}=K(x,y)\frac{\partial^2 F}{\partial z^2}, \tag{4.14}$$

where $K\neq 0$. But then (4.13) implies that

$$c^2=1. \tag{4.15}$$

Integrate (4.14),

$$\frac{\partial^3 F}{\partial z^3}=K\left(\frac{\partial F}{\partial z}+K_1(x,y)\right). \tag{4.16}$$

Multiply (4.16) by $2\partial^2 F/\partial z^2$ and integrate, i.e.,

$$\left(\frac{\partial^2 F}{\partial z^2}\right)^2 = K\left[\left(\frac{\partial F}{\partial z}\right)^2 + 2K_1\frac{\partial F}{\partial z}\right] + K_2. \tag{4.17}$$

Substitute (4.16) and (4.17) into (4.12),

$$2KK_1\left[\frac{\partial F}{\partial w} - \frac{\partial F}{\partial z}\right] = 0.$$

But F is not linear,

$$\therefore\ K_1 = 0. \tag{4.18}$$

(4.16) and (4.18) then give the result

$$\frac{\partial^3 F}{\partial z^3} = K\frac{\partial F}{\partial z}. \tag{4.19}$$

5. CONCLUSION

Equation (4.15) gives $c = \pm 1$, so suppose

$$c = +1. \tag{5.1}$$

Define a function $\rho(x,y)$ by

$$[\rho(x,y)]^2 = K(x,y). \tag{5.2}$$

Note that ρ is either real or pure imaginary. So for this section all constants of integration may be complex. At the end, one will demand that F, P_0 and Q_0 are real.

The general solution of (4.19) is

$$\frac{\partial F}{\partial z} = A_1(x,y)\exp(\rho z) + A_2(x,y)\exp(-\rho z). \tag{5.3}$$

Substitute (5.3) into (4.2),

$$2\frac{\partial P_0}{\partial w}\frac{\partial Q_0}{\partial w} = (A_1\exp[\rho(w+z)/2] + A_2\exp[-\rho(w+z)/2]) \times (\exp[\rho(w-z)/2] + \exp[\rho(w-z)/2]). \tag{5.4}$$

Now (2.11) is $\partial P_0/\partial w = \partial P_0/\partial z$, so P_0 is a function of $w+z$ only and similarly, from (2.14), Q_0 is a function of $w-z$ only.

Hence (5.4) implies for some function $\lambda(x,y) \neq 0$:

$$\frac{\partial P_0}{\partial w} = \lambda(x,y)[A_1\exp(\rho(w+z)/2) + A_2\exp(-\rho(w+z)/2)], \tag{5.5}$$

$$\frac{\partial Q_0}{\partial w} = \frac{1}{2\lambda}[\exp(w-z)/2) + \exp(-\rho(w-z)/2)]. \tag{5.6}$$

Integrating (5.3), (5.5), and (5.6) and using the fact that P_0 and Q_0 are functions of $w+z$ and $w-z$ respectively one obtains

$$F(x,y,z) = \frac{A_1}{\rho}\exp(\rho z) - \frac{A_2}{\rho}\exp(-\rho z) + K_1(x,y), \tag{5.7}$$

$$P_0(x,y,z,w) = (2\lambda A_1/\rho)\exp[\rho(w+z)/2] - (2\lambda A_2/\rho)\exp[-\rho(w+z)/2] + K_2(x,y) \tag{5.8}$$

$$Q_0(x,y,z,w) = (1/\lambda\rho)\exp[\rho(w-z)/2] - (1/\lambda\rho)\exp[-\rho(w-z)/2] + K_3(x,y). \tag{5.9}$$

Substitute (5.7), (5.8), and (5.9) into (2.12) and (2.16) and equate coefficients of $\exp[\pm(1/2)\rho z]$ and terms independent of $\exp[\pm(1/2)\rho z]$.

$$\lambda A_1 K_3 + \frac{\partial}{\partial y}\left(\frac{2\lambda A_1}{\rho}\right) + \frac{\lambda A_1}{\rho}(w+z)\frac{\partial \rho}{\partial y} = 0, \tag{5.10}$$

$$\lambda A_2 K_3 - \frac{\partial}{\partial y}\left(\frac{2\lambda A_2}{\rho}\right) + \frac{\lambda A_2}{\rho}(w+z)\frac{\partial \rho}{\partial y} = 0, \tag{5.11}$$

$$\frac{\partial K_2}{\partial y} = 0, \tag{5.12}$$

$$\frac{K_2}{2\lambda} + \frac{\partial}{\partial x}\left(\frac{1}{\lambda\rho}\right) + \frac{1}{2\lambda\rho}\frac{\partial \rho}{\partial x}(w-z) = 0, \tag{5.13}$$

$$\frac{K_2}{2\lambda} - \frac{\partial}{\partial x}\left(\frac{1}{\lambda\rho}\right) + \frac{1}{2\lambda\rho}\frac{\partial \rho}{\partial x}(w-z) = 0, \tag{5.14}$$

$$\frac{\partial K_3}{\partial x} - 2K_1 = 0. \tag{5.15}$$

The coefficient of $(w+z)$ in (5.10) and (5.11) must be zero and one cannot have both $A_1 = 0$ and $A_2 = 0$. Hence $\partial\rho/\partial y = 0$. The coefficient of $(w-z)$ in (5.13) is zero so $\partial\rho/\partial x = 0$

$$\therefore\ \rho \text{ is a constant.} \tag{5.16}$$

Add and subtract (5.13) and (5.14) and use (5.16), i.e.,

$$K_2 = 0, \tag{5.17}$$

$$\frac{\partial \lambda}{\partial x} = 0. \tag{5.18}$$

Let $\alpha(x,y)$ be any function such that

$$K_3 = \frac{\partial \alpha}{\partial y}. \tag{5.19}$$

Using (5.16) and (5.19), Eqs. (5.10) and (5.11) are

$$\frac{\partial}{\partial y}(\lambda A_1) + \lambda A_1\frac{\partial}{\partial y}(\tfrac{1}{2}\rho\alpha) = 0, \tag{5.20}$$

$$\frac{\partial}{\partial y}(\lambda A_2) - \lambda A_2\frac{\partial}{\partial y}(\tfrac{1}{2}\rho\alpha) = 0. \tag{5.21}$$

Integrating (5.20) and (5.21) gives

$$\lambda A_1 = \beta(x)\exp(-\rho\alpha/2), \tag{5.22}$$

$$\lambda A_2 = \gamma(x)\exp(\rho\alpha/2). \tag{5.23}$$

Equation (5.15) gives

$$K_1 = \frac{1}{2}\frac{\partial K_3}{\partial x} = \frac{1}{2}\frac{\partial^2 K}{\partial x \partial y}. \tag{5.24}$$

From Eqs. (5.7)–(5.9), (5.16)–(5.19), and (5.22)–(5.24) one has, for $c = +1$, the following result: The equation $z_{xy} = F(x, yz)$ possesses a Bäcklund transformation if and only if

$$F(x,y,z) = (1/\lambda(y)\rho)[\beta(x)\exp[\rho(z-\alpha/2)] - \gamma(x)\exp[-\rho(z-\alpha/2)]] + \frac{1}{2}\frac{\partial^2 \alpha}{\partial x \partial y}. \tag{5.25}$$

The Backlund transformation is then

$$\frac{\partial w}{\partial x} = [2\beta(x)/\rho]\exp[\rho(w+z-\alpha)/2] - [2\gamma(x)/\rho]\exp[\rho(w+z-\alpha)/2] + \frac{\partial z}{\partial x}, \tag{5.26}$$

$$\frac{\partial w}{\partial y} = [1/\rho\lambda(y)]\exp[\rho(w-z)/2] - [1/\rho\lambda(y)]$$

$$\times\exp[-\rho(w-z)/2] + \frac{\partial \alpha}{\partial y} - \frac{\partial z}{\partial y}\,. \qquad (5.27)$$

The result for $c=-1$ is easily obtained by making the substitutions $c \leftrightarrow -c$, $P_0 \leftrightarrow Q_0$, $x \leftrightarrow y$ for then with $P_1 = -Q_1 = c$, the Eqs. (2.9)—(2.16) transform into themselves.

Now it is always possible to find a real function, $H(z)$, of a single variable which satisfies $H''(z) = \rho^2 H(z)$ where ρ^2 is real and satisfies for some functions $R(x)$ and $\epsilon(x)$

$$F(x,y,z) = \frac{R(x)}{\lambda(x)} H[z - \tfrac{1}{2}\alpha(x,y) + \epsilon(x)] + \frac{1}{2}\frac{\partial^2 \alpha}{\partial x \partial y},$$

where F is given by Eq. (5.25). But clearly if one displaces z by an amount $\frac{1}{2}\alpha - \epsilon$ and also changes the scale, then the equation $z_{xy} = F(x,y,z)$ becomes $z_{xy} = H(z)$.

So it has been proved that if the equation $z_{xy} = F(x,y,z)$ possesses a Bäcklund transformation which takes solutions into other solutions of the same equation then the equation must be $z_{xy} = F(z)$ where $F(z)$ satisfies $F''(z) = KF(z)$ for a real constant $K \neq 0$ (up to a scale change and a displacement in z).

ACKNOWLEDGMENTS

I wish to thank Dr. D.B. Fairlie for suggesting this problem and for helpful discussions. I would also like to thank the British Council for providing my research grant.

[1]For a general review of solitons, see, for example, A.C. Scott, F.Y.F. Chu, and D.W. McLaughlin, Proc. IEEE **61**, 1443 (1973); G.W. Whitham, *Linear and Non-Linear Waves* (Wiley, New York, 1974).

[2]D.W. McLaughlin and A.C. Scott, J. Math. Phys. 14, 1817 (1973).

[3]R. Hirota, J. Phys. Soc. Jpn. 35, 1566 (1973).

[4]S. Coleman, Phys. Rev. D **11**, 2088 (1975). S. Mandlestam, University of California, Preprint.

[5]G.L. Lamb Jr., J. Math. Phys. **15**, 2157 (1974).

[6]R.K. Dodd, R.K. Bullough, and S. Duckworth [J. Phys. A 8, L64 (July, 1975)] found that the equation $\sigma_{xx} - \sigma_{tt} = \sin\sigma + \frac{1}{2}\sin\frac{1}{2}\sigma$ has "a wealth of unusual solutions of multisoliton type." These were found by computer studies and suggest that there are equations, of second order, which possess the soliton property but do not have Bäcklund transformations.

[7]I wish to thank the referee of this paper for noting that in the analysis of the "quantum flux shuttle" [T.A. Fulton, R.C. Dynes, and P.W. Anderson, Proc. IEEE 61, 28 (1973)] the result obtained here is of use. In the one-dimensional, no-damping quantum flux shuttle the equation

$$\phi_{xx} - \phi_{tt} = a(x)\sin\phi\,,$$

$$a(x) = a(x + x_0)\,,$$

occurs. In characteristic coordinates this is

$$\phi_{xy} = a(x+y)\sin\phi\,.$$

So one sees that for real ϕ, this equation only has a Bäcklund transformation if a is a constant.

Series of Stieltjes, Padé approximants and continued fractions*

David A. Field

College of the Holy Cross, Worcester, Massachusetts
(Received 2 September 1975)

Nested sequences of lens-shaped regions which contain Padé and continued fraction approximants to series of Stieltjes are investigated. It is shown by continued fraction methods that the recent results on Padé approximants by Baker and results on continued fractions by Gragg yield identical sequences.

Considerable attention has recently been given to Padé and continued fraction approximants.[1-7] Of particular interest are the results of Baker[7] and Gragg[8] who have defined apparently different nested sequences of lens-shaped regions which contain Padé and continued fraction approximants of series of Stieltjes. We will show via continued fraction methods that the sequences are identical and hence the criteria defining these sequences are equivalent.

Baker investigated series of Stieltjes, $f(z)$,

$$f(z)=\sum_{i=0}^{\infty}\alpha_i\,(-z)^i \tag{1}$$

with radius of convergence $R>0$ and with continued fraction representations of the form

$$f(z)=\frac{a_0}{1+}\,\frac{a_1z}{1+}\,\frac{a_2z}{1+}\,\cdots, \tag{2}$$

where a_n, $n\geq 0$, is positive and is defined by the relations

$$f_0(z)=f(z),\quad f_n(z)=\frac{a_n}{1+zf_{n+1}(z)},$$

$f_n(z)\neq 0$, and $f_n(0)\neq 0$, $n\geq 0$.

Gragg examined analytic functions, holomorphic for $|\arg(1+z)|<\pi$, satisfying $\mathrm{Re}\{[1+zf(z)]^{1/2}\}>0$, the principal branch of the square root being assumed in this domain, and which have continued fraction representations of the form

$$f(z)=\frac{g_0}{1+}\,\frac{g_1z}{1+}\,\frac{(1-g_1)g_2z}{1+}\,\frac{(1-g_2)g_3z}{1+}\,\cdots, \tag{3}$$

where

$$|\arg(1+z)|<\pi,\quad g_0>0 \text{ and } 0<g_n<1,\quad n\geq 1.$$

That Baker and Gragg are considering the same analytic functions can be seen as follows.

$f(z)$ is a series of Stieltjes with radius of convergence R if and only if there exists a bounded nondecreasing function σ with infinitely many points of increase on $[0,\ 1/R]$ such that

$$f(z)=\int_0^{1/R}\frac{d\sigma(t)}{1+tz},\quad |\arg(R+z)|<\pi. \tag{4}$$

Since changes in variables $\zeta=z/R$ and $\tau=Rt$ normalize (4), we can assume that $R=1$. Wall[9] proved that $f(z)$ is a series of Stieltjes with radius of convergence $R=1$ if and only if $f(z)$ has a continued fraction representation of the form (3). Furthermore, it can be shown with a proof analogous to Leighton and Scott's theorem[10] that the continued fractions in (2) and (3) are identical.

It will be essential for us to consider continued fractions in terms of linear fractional transformations. If s_n denotes the linear fractional transformation

$$s_n(w)=\frac{a_{n-1}z}{1+w},\quad n>1,\quad s_1(w)=\frac{a_0}{1+w} \tag{5}$$

and

$$S_1(w)=s_1(w),\quad S_n(w)=S_{n-1}(s_n(w)),\quad n\geq 2, \tag{6}$$

then the nth approximant of the continued fraction (3) is

$$S_n(0)=\frac{A_n(z)}{B_n(z)}=\frac{a_0}{1+}\,\frac{a_1z}{1+}\,\frac{a_2z}{1+}\,\cdots\frac{a_{n-1}z}{1}, \tag{7}$$

where A_n and B_n satisfy the relations

$$A_n(z)=a_{n-1}z\,A_{n-2}(z)+A_{n-1}(z),$$
$$A_0(z)=0,\quad A_1(z)=a_0,\quad n\geq 2, \tag{8}$$

and

$$B_n(z)=a_{n-1}z\,B_{n-2}(z)+B_{n-1}(z),$$
$$B_1(z)=B_0(z)=1,\quad n\geq 2. \tag{9}$$

We state our main theorem.

Theorem: If $f(z)$ is a series of Stieltjes with radius of convergence $R>0$ with power series and continued fraction representations (1) and (3), then $\{H(a_0,\ldots,a_{n-1};z)\}$, $n\geq 1$, a nested sequence of regions each containing $f(z)$, can be defined in four equivalent ways.

H1. For each $w\in H_n(a_0,\ldots,a_{n-1};z)$ there exists a series of Stieltjes $f^*(z)$ with radius of convergence R whose first n terms agree with $f(z)$ and $f^*(z)=w$.

H2. $H_n(a_0,\ldots,a_{n-1};z)$, $n\geq 1$ is the interior of a region whose boundary consists of two circular arcs $\gamma_n(z)$ and $\Gamma_n(z)$ defined by

$$\gamma_n(z)=\{S_n(c_nzw):\ 0\leq w\leq 1/R\} \tag{10}$$

$$\Gamma_n(z)=\left\{S_n(c_nzw):\ w=\frac{1-uR}{R(1+uz)},\ 0\leq u\leq 1/R\right\} \tag{11}$$

where

$$c_n=B_n(-R)/B_{n-1}(-R).$$

H3. $H_n(a_0,\ldots,a_{n-1};z)-H_{n+1}(a_0,\ldots,a_{n-1},a_n;z)$, $n\geq 1$ consists of two components $L_n(z)$ and $L_n'(z)$. $L_n(z)$ $[L_n'(z)]$ is a circular triangle with vertices $w_{n-1}(z)$, $w_n(z)$, $w_n'(z)$, $[w_{n-1}'(z)$, $w_n'(z)$, $w_n(z)]$ and respective interior angles $\theta=|\arg(1+z)|$, $\phi=|\arg[z/(1+z)]|$, $\psi=|\arg(-1/z)|$ where $w_0'(z)$, $w_0(z)=S_1(0)$, $w_1'(z)$; $w_1(z)=S_2(0)$, $w_2'(z),\ldots$ are the successive approximants of the continued fraction

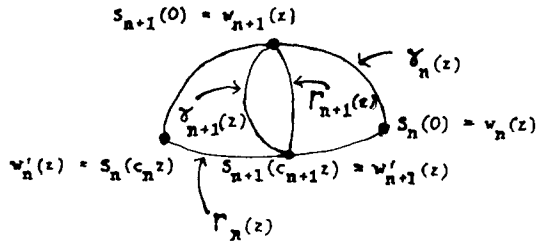

FIG. 1

$$\frac{\pi_0}{1+z+}\ \frac{-z}{1+}\ \frac{\pi_1}{1+z+}\ \frac{-z}{1+}\ \cdots,$$

$$\pi_0 = g_0, \quad \pi_n = g_n/(1-g_n), \quad n \geq 1. \tag{12}$$

H4. The open convex region $H_n(a_0, \ldots, a_{n-1}; z)$, $n \geq 1$, is defined by

$$H_n(a_0, \ldots, a_{n-1}; z) = \{v : v = w'_{n+2}(z) \text{ or } v = w_{n+2}(z)\},$$

where $w'_n(z)$, $w_n(z)$ are defined in H3. Furthermore, the diameter of $H_n(a_0, \ldots, a_{n-1}; z)$ is at most

$$\max\{1, \tan\theta/2\}\ |S_n(c_n z) - S_n(0)|. \tag{13}$$

Proof: We begin our proof with a few important observations. The interior angles of the circular triangles $L_n(z)$ and $L'_n(z)$ imply that the points $w_{n-1}(z)$, $w_n(z)$, $w'_{n-1}(z)$ and the points $w'_{n-1}(z)$, $w'_n(z)$, $w_{n-1}(z)$ define circular arcs $\kappa_n(z)$ and $\kappa'_n(z)$, respectively. Gragg[11] observed that $\kappa_n(z)$ and its circular extension pass through $w_n(z)$, $w_{n-1}(z)$, and $w_{n-2}(z)$ which are $S_{n+1}(0)$, $S_n(0)$, and $S_{n-1}(0)$ respectively. Finally, since Gragg utilized Wall's result and the normalization of (4), we will assume $R = 1$.

Baker[12] proved the equivalence of H1 and H2 while the equivalence of H3 and H4 is stated in Gragg's Theorem 3.[13] We will prove the equivalence of H2 and H3 by proving that $\gamma_n(z) = \kappa_n(z)$ and $\Gamma_n(z) = \kappa'_n(z)$.

To prove $\gamma_n(z) = \kappa_n(z)$ it suffices to demonstrate that the circular arcs $\gamma_n(z)$ and $\kappa_n(z)$ and their respective circular extensions have three common points. $\gamma_n(z)$ and its circular extension is defined by the set $\{w : w = S_n(c_n z u),\ -\infty \leq u \leq \infty\}$ which contains the points $S_n(0)$, $S_{n-1}(0)$, and $S_{n+1}(0)$ for the values $u = 0, \infty$, and a_n/c_n respectively. Since $\kappa_n(z)$ and its circular extension passes through $S_n(0)$, $S_{n-1}(0)$ and $S_{n+1}(0)$, we have $\gamma_n(z) = \kappa_n(z)$.

To prove $\Gamma_n(z) = \kappa'_n(z)$, we begin with the fact that $S_n(w)$ is a conformal mapping and $|\arg(1+z)|$ is the interior angle between the curves

$$\gamma = \{w : 0 \leq W \leq 1\} \text{ and } \Gamma = \left\{ w : w = \frac{1-u}{1+uz},\ 0 \leq u \leq 1 \right\}.$$

Thus the angle between the curves $\gamma_n(z)$ and $\Gamma_n(z)$ and the curves $\kappa_n(z)$ and $\kappa'_n(z)$ is $|\arg(1+z)|$. Furthermore, $\Gamma_n(z)$ with its circular extension is the unique circle intersecting $\gamma_n(z)$ at $S_n(0)$ with angle $|\arg(1+z)|$ and passing through $S_n(c_n z)$, and $\kappa'_n(z)$ with its circular extension is the unique circle which intersects $\gamma_n(z)$ at $S_n(0)$ with angle $|\arg(1+z)|$ and passing through $w'_n(z)$. Hence to show $\kappa'_n(z) = \Gamma_n(z)$, it will be sufficient to prove via induction

$$w'_n(z) = S_n(c_n z). \tag{14}$$

First, we show $a_{n+1} < c_n$ from which we conclude that $S_{n+1}(c_{n+1}z)$ lies on $\Gamma_n(z)$ between $S_n(0) = S_n(c_n z)$. Stieltjes[14] proved that the roots of $B_n(z)$ lie on the interval $(-\infty, -R)$. Furthermore it is an immediate consequence from (9) that the leading coefficient of $B_n(z)$ is positive. Thus our assumption that $R = 1$ and division of (9) by $B_n(-1)$ yields

$$0 < c_n = 1 - a_n/c_{n-1} \quad \text{or} \quad c_{n-1} > a_n \text{ for } n \geq 2. \tag{15}$$

We now assert the existence of u', $0 \leq u' \leq 1$ such that

$$S_n\left(\frac{c_n z(1-u')}{1+u'z}\right) = S_n \frac{a_{n+1}z}{1+u'z}. \tag{16}$$

Since the above equation equates two values of a linear fractional transformation which is a one-to-one function we solve this equation for u' to obtain $u' = c_{n+1}$. Consequently, from (6), $S_n(s_{n+1}(c_{n+1}z)) = S_{n+1}(c_{n+1}z)$ and $S_{n+1}(c_{n+1}z)$ lies on $\Gamma_n(z)$ between $S_n(0)$ and $S_n(c_n z)$.

Our induction begins with the fact that (2) and (3) are identical continued fraction representations so that in (12), $\pi_1 = a_1/(1-a_1)$. It is then easily verified using Eqs. (6), (8) and (9) that

$$w'_1(z) = a_0/(1+z) = S_1(c_1 z)$$

and

$$w'_2(z) = \frac{a_0}{1+z+}\ \frac{-z}{1+}\ \frac{\pi_1}{1+z}$$

$$= \frac{a_0 + a_0 c_1 z}{1 + c_1 z + a_1 z} = S_2(c_2 z).$$

Therefore $\Gamma_1(z) = \kappa'_1(z)$ and $\Gamma_2(z) = \kappa'_2(z)$ and we assume the induction hypothesis, $\Gamma_n(z) = \kappa'_n(z)$. Since $S_{n+1}(c_{n+1}z)$ also lies on $\kappa_n(z) = \Gamma_n(z)$, and since $\Gamma_n(z)$ and $\gamma_{n+1}(z)$ and their circular extensions can have only two points in common one of which is $S_n(0)$, we have $S_{n+1}(c_{n+1}z) = w'_{n+1}(z)$ and therefore $\Gamma_{n+1}(z) = \kappa'_{n+1}(z)$.

An elementary geometric argument shows that the diameter of $H_n(a_0, \ldots, a_n; z)$ is at most equal to the expression in (13) and our theorem is proved.

*This research was supported by the United States Air Force through the Air Force Office of Scientific Research under Grant No. AFOSR-70-1888.

[1]G.A. Baker, Jr., J. Math. Phys. **10**, 814 (1969).
[2]A.K. Common, J. Math. Phys. **9**, 32 (1968).
[3]R.G. Gordon, J. Math. Phys. **9**, 1087 (1968).
[4]W.B. Gragg, Numer. Math. **11**, 370 (1968).
[5]P. Henrici and P. Pfluger, Numer. Math. **9**, 120 (1966).
[6]W.B. Jones and W.J. Thron, SIAM J. Numer. Anal. **8**, 693 (1971).
[7]See Ref. 1, pp. 815–16.
[8]See Ref. 4, Theorem 3.
[9]H.S. Wall, *Analytic Theory of Continued Fractions* (Van Nostrand, New York, 1948).
[10]W. Leighton and W.T. Scott, Bull. Am. Math. Soc. **45**, 596 (1939).
[11]See Ref. 4, Eqs. (14)–(16).
[12]See Ref. 1, pp. 815–16.
[13]See Ref. 4.
[14]T.J. Stieltjes, *Oeuvres* (Noordhoff, Groningen, 1914-18), Vol. 2, pp. 402–566.

The smooth-path topology for curved space–time which incorporates the conformal structure and analytic Feynman tracks

Rüdiger Göbel

Universität Essen, GHS, Fachbereich 6, Mathematik, D 4300 Essen, Germany
(Received 18 July 1975)

It is obvious that the usual (*Riemannian*) topology on a space–time has no natural relation with the (*pseudo-Riemannian*) metric of general (or special) relativity. Therefore, several new topologies on a space–time were proposed in recent years in order to overcome this "classical" disharmony. In this paper an infinite variety of more physical topologies is investigated (including the most interesting known topologies). It turns out that one of the candidates "$\mathfrak{P}_\omega^*$" has all desired physical advantages: (1) It is defined in a very physical way by tests with particles of mass >0. (2) It carries intrinsically all information about the smooth conformal structure of space–time: Its homeomorphism group is the conformal group. (3) The (only) 1–1 continuous curves are the analytic Feynman tracks. It turns out that this topology $\mathfrak{P}_\omega^*$ is strictly finer than the topology suggested by S. W. Hawking, A. R. King, and P. J. McCarthy. During this investigation a conjecture of E. C. Zeeman (1967; for Minkowski space) will be proved for all strongly causal space–times (including Minkowski space).

1. INTRODUCTION

In 1967 Zeeman[1] published his paper suggesting a new topology (here denoted by) $\mathfrak{Z}_g$ on Minkowski space which is defined more physically than the ordinary (Euclidian) topology $\mathfrak{T}$. He suggested generalizing this topology $\mathfrak{Z}_g$ to general relativity and deriving the corresponding results for curved space–times. This was worked out by R. Göbel.[2] It turns out that on a curved space–time $\mathfrak{Z}_g$ has many physical properties, for instance (A), (B), and (C).

(A) *It can be "tested" partly* [= condition (*)] *by particles* in the sense of thought experiments:

A subset **X** of the space–time **M** belongs to $\mathfrak{Z}_g$ if and only if

(*) $\mathbf{X}\cap g$ is open in g (in the sense of the usual topology induced by the eigentime of an observer on g) for all world lines g of freely falling test particles of mass >0;

(**) $\mathbf{X}\cap\mathbf{Y}$ carries the Euclidian topology for all spacelike hypersurfaces **Y** of **M**.

A map $f:[0,1]\to\mathbf{M}$ which is 1–1 and continuous with respect to the underlying topology is called a *path* and its image $f=f[0,1]$ the corresponding *curve*. Then f is a C^n-(smooth)curve if f is an at least n-times (continuously) differentiable map. f will be called *broken* (C^n-*path/curve*) if there are finitely many exceptional points only, where f may have two tangents. It is common to call f a *world line*[2] [called timelike path in Hawking, King, and McCarthy[3] (HKM)] if f is a ($\mathfrak{T}$-) continuous path which is order preserving, i. e.,

if $a\in[0,1]$, there is a neighborhood **U** of a in $[0,1]$ and a causal (simple convex) neighborhood **V** of $f(a)$ (cf. Ref. 4, p. 5) such that from $b<c$ in **U** follows $f(b)\ll f(c)$. Here $x\ll y$ means that x belongs to the interior of the past light cone of y within **V**; cf. Ref. 4, p. 11.

(B) *Continuous world lines with respect to* $\mathfrak{Z}_g$ are now no longer mathematical (pathological) curves: They *are broken timelike, future directed geodesics*, i. e., tracks of freely falling test particles within the gravitational field with finitely many bounces (as used in kinetic theory); cf. Göbel[2] (Corollary 3. 6).

(C) The group of homeomorphisms of an *arbitrary* space–time **M** with respect to $\mathfrak{Z}_g$ (is no longer a vast set and) turns out to coincide with the groups of all homothetic transformations of **M** (they equal isometries up to a universal constant); cf. Ref. 2 (Corollary 5. 11). In the case of Minkowski space this is the group generated by Lorentz transformations and (constant) dilatations; cf. Zeeman[1] (p. 168, Theorem 3).

Besides these attractive features, $\mathfrak{Z}_g$ has some physical disadvantages, as pointed out in HKM,[3] i. e.,

(1*) The property (A) is *only partly physical*; condition (**) should be removed.

(2*) The group of homothetic transformations seems to be physically less important than the conformal group. Therefore it would be more satisfactory to obtain the conformal group or the isometry group as the invariant group of the topology. [However for positive remarks about the physical importance of homothetic transformations we refer you to Einstein[5] and for further discussions and references to Winicour.[6]]

(3*) World lines of particles which are accelerated under forces different from the gravitational forces are no longer continuous.

(4*) $\mathfrak{Z}_g$ is technically complicated.

Criticism (3*) was removed in Ref. 2 at least for charged particles within an external electromagnetic field, considering all "*e/m-orbits*". Because of (1*), (2*), and partly (3*) and (4*), a new topology $\mathfrak{P}$ (here called $\mathfrak{P}_0$) was introduced in HKM.[3] $\mathfrak{P}_0$ has many physically attractive properties. As a result of this

the objections (1*), (2*), (3*), and (partly) (4*) do no longer exist.

The topology $\mathfrak{P}_0$ [derived as a result of criticisms (1*)–(4*)], however, is liable to the following objections:

(1**) The topology $\mathfrak{P}_0$ is defined by (A) (*) for *all* timelike C^0-paths g. Therefore it is difficult to think of a "test" for $\mathfrak{P}_0$ (in the sense of a thought experiment) by particles, since there are many timelike C^0-paths which have no physical meaning as tracks of certain particles, e.g., *"bad trips"* in in the sense of Penrose[4] (p. 11).

(3**) Among the possible $\mathfrak{P}_0$-continuous curves and world lines there are many unphysical ones with a high degree of indifferentiability; cf. HKM[3] (Theorem 2). It would be nice to be able to single them out by the continuity requirement only!

Therefore we ask, whether it is possible to refine this topology $\mathfrak{P}_0$, suggested by Hawking (cf. Ref. 2, p. 297) in the topological and the real sense—and to balance a new topology $\mathfrak{X}$ in such a way that $\mathfrak{X}$ has all additional nice physical properties of $\mathfrak{P}_0$, and that the criticisms (1**) and (3**) can be overcome.

In order to have the freedom of choice for such a topology we will introduce four infinite sequences (!) of topologies and will pick our best candidate $\mathfrak{P}^*_\omega$ later on, i.e.,

$$(+)\quad \begin{array}{ccccccccccccccc} \mathfrak{P}^*_0 & = & \mathfrak{P}^*_1 & < \cdots < & \mathfrak{P}^*_n & < & \mathfrak{P}^*_{n+1} & < \cdots < & \mathfrak{P}^*_\infty & < & \mathfrak{P}^*_\omega & < & \mathfrak{P}^*_g = \mathfrak{P}_g, \\ \vee & & \vee & & \vee & & \vee & & \vee & & \vee & & \| \\ \mathfrak{P}_0 & = & \mathfrak{P}_1 & < \cdots < & \mathfrak{P}_n & < & \mathfrak{P}_{n+1} & < \cdots < & \mathfrak{P}_\infty & < & \mathfrak{P}_\omega & < & \mathfrak{P}_g, \\ \vee & & \vee & & \vee & & \vee & & \vee & & \vee & & \vee \\ \mathfrak{Z}_0 & = & \mathfrak{Z}_1 & < \cdots < & \mathfrak{Z}_n & < & \mathfrak{Z}_{n+1} & < \cdots < & \mathfrak{Z}_\infty & < & \mathfrak{Z}_\omega & < & \mathfrak{Z}_g, \\ \wedge & & \wedge & & \wedge & & \wedge & & \wedge & & \wedge & & \wedge \\ \mathfrak{Z}^*_0 & = & \mathfrak{Z}^*_1 & < \cdots < & \mathfrak{Z}^*_n & < & \mathfrak{Z}^*_{n+1} & < \cdots < & \mathfrak{Z}^*_\infty & < & \mathfrak{Z}^*_\omega & < & \mathfrak{Z}^*_g = \mathfrak{Z}_g. \end{array}$$

The topologies $\mathfrak{Z}_g$ and $\mathfrak{P}_g$ investigated in Göbel[2] are on the very right, and the topology $\mathfrak{P}_0$ suggested in HKM[3] is on the very left of the diagram (+). The reason for the introduction of the last two sequences in (+) and $\mathfrak{P}_g$ *is to answer a further question of* Zeeman[1] (the conjecture on p. 169 in Ref. 1 for Minkowski space, which will be proved for all strongly causal space–times); cf. Corollary 6.3. Furthermore the $\mathfrak{Z}$-types of topologies might be more useful in connection with initial data problems, since space like hypersurfaces are still endowed with the common Euclidean topology.

The physically most attractive topology seems to be one of the $\mathfrak{P}$-topologies. In order to define $\mathfrak{P}^*_k$, we need the following notation: We call a world line f *strictly timelike at* $p \in f$, if no sequence of geodesics $\overline{px}$ (through p and x within a simple neighborhood of p) converges to the null cone if $x \in f$ and $x \to p$ with respect to $\mathfrak{T}$. The topology on a timelike path can be found in two equivalent ways, either as the topology induced by $\mathfrak{T}$ or induced by the metric of the space–time. With this topology on the timelike paths we define for a subset $\mathbf{X}$ of the space–time $\mathbf{M}$,

(++) $\mathbf{X} \in \mathfrak{P}_n$ ($\mathbf{X} \in \mathfrak{P}^*_n$) if and only if $\mathbf{X} \cap g$ is open in g for all (strictly timelike) C^n-world lines g of $\mathbf{M}$.

It is always assumed that the degree of differentiability of the manifold $\mathbf{M}$ is at least n. If $n = \infty$ this means that $\mathbf{M}$ and the world-lines g in (++) are arbitrarily many times differentiable, and if $n = \omega$, $\mathbf{M}$ and g in (++) are analytic. If $n = g$, the world lines g in (++) are timelike geodesics, $\mathbf{M}$ is at least three-times differentiable and the metric is C^2.

(+++) If condition (A) (**) is fulfilled in conjunction with (++) for C^n-world lines, we obtain $\mathbf{X} \in \mathfrak{Z}_n$ or $\mathbf{X} \in \mathfrak{Z}^*_n$.

It can be seen from the very definition, that objection (1**) is removed in the case (++) if n is sufficiently large. Surprisingly $\mathfrak{P}_0 = \mathfrak{P}_1$ and $\mathfrak{P}^*_0 = \mathfrak{P}^*_1$ as well as $\mathfrak{Z}_0 = \mathfrak{Z}_1$ and $\mathfrak{Z}^*_0 = \mathfrak{Z}^*_1$ coincide, which means that the topology $\mathfrak{P}_0$ in HKM[3] can be defined by the (less wild) C^1-paths. It can be shown that all other topologies of our diagram (+) do not collapse; cf. Lemma 3.2. Therefore it might be possible to find criteria to select between, for instance $\mathfrak{P}_n$ and $\mathfrak{P}_{n+1}$. The essential difference between the $*$-topologies and the topologies without a "$*$" is that in the latter case world lines (i.e., with respect to the non-"$*$"-topology) may converge to light directions, which is excluded in the $*$-type.

Criticism (4*) is a question of tradition. In particular I believe that *it is simpler to calculate in topology if there are very many or very few open sets*. The first case arises at the right-end of the sequences of our diagram (+).

In addition it can be shown that the homeomorphism groups in all cases not equal to $\mathfrak{P}_g$ nor $\mathfrak{Z}_g$ are the conformal groups; cf. Theorems 6.2 and 6.4. For the $\mathfrak{Z}$-type topologies this is true even without the restriction "strongly causal" on $\mathbf{M}$. *We conjecture* that this restriction is not necessary in the other cases either. In the case $\mathfrak{P}_g$, the homeomorphisms are the homothetic transformations as in the case $\mathfrak{Z}_g$; cf. Corollary 6.3. *This* finally *answers the question by Zeeman*, mentioned above; cf. Nanda[7] and Geroch.[8]

Next we concentrate on the *"analytic topologies"* with $n = \omega$. The criticism (3*) and the corresponding criticism (3**) are removed! We call a broken analytic timelike path which is piecewise past or future directed *an analytic Feynman track*. These are special Feynman tracks as defined in HKM.[3] Then $\mathfrak{P}_\omega$-continuous paths are *analytic* Feynman tracks and $\mathfrak{P}^*_\omega$-continuous paths are in addition strictly timelike. Therefore world-lines with respect to $\mathfrak{P}_\omega$ (or $\mathfrak{P}^*_\omega$) are broken analytic (strictly) timelike and future directed paths. The same holds for world lines with respect to $\mathfrak{Z}_\omega$ and $\mathfrak{Z}^*_\omega$; cf. Secs. 4 and 5. Due to the natural definition and the physical properties *we would like to recommend* (like S. W. Hawking, A. R. King, and P. J. McCarthy) one of the $\mathfrak{P}$-type topologies, but in particular $\mathfrak{P}_\omega$ or $\mathfrak{P}^*_\omega$! As long as we allow analytic or broken analytic curves of particles of rest mass > 0 for tests of the topology of space–time, we obtain $\mathfrak{P}^*_\omega$ as the natural topology. It might, however, be interesting for mathematical reasons and for a better understanding of $\mathfrak{P}^*_\omega$ to look at the world lines with respect to $\mathfrak{P}_n$ ($n \neq 0, 1, \omega, g$), too. The results expected are somehow "between $\mathfrak{P}_0$ and $\mathfrak{P}_\omega$."

The results in this paper are derived by application of the beautiful ideas and results in HKM[3] and by some results in Göbel.[2] We will use the notation in Refs. 2 and 3.

2. MATHEMATICAL TOOLS

In order to investigate Zeeman topologies on a curved space–time we need the following propositions. In particular, Proposition 2.1 will be necessary for redefining the topology $\mathfrak{P}_0$ suggested by Hawking, cf. Ref. 2 and HKM,[3] with the help of C^1-world-lines. Proposition 2.4 is used to derive physical properties of the Zeeman topology $\mathfrak{P}^*_\omega$ suggested in Sec. 1.

Proposition 2.1: Let **M** be the Minkowski space and **S** a sequence of points $p_n \in \mathbf{M}$ labelled by the natural numbers with the following properties:

(1) $\mathbf{S} \to p$, i.e., there is a point $p \in \mathbf{M}$ such that **S** converges to p with respect to the ordinary (Euclidian) topology $\mathfrak{T}$ on **M**.

(2) If $x \in \mathbf{S}$, then $p \neq x$ are timelike related in **M**.

There is a C^1-world-line passing through p and infinitely many points of **S**. This world line is timelike at each point $\neq p$ and timelike or lightlike at p.

Remark: C^1 is *"best possible"* as follows from Proposition 2.3. *Construction of the required C^1-world-line*: Because of condition (2) we get $\mathbf{S} \subset \mathbf{I}^-(p) \cup \mathbf{I}^+(p)$. Therefore $\mathbf{S} \cap \mathbf{I}^-(p)$ or $\mathbf{S} \cap \mathbf{I}^+(p)$ is infinite. Let us assume, without loss of generality, $\mathbf{S}_1 = \mathbf{S} \cap \mathbf{I}^+(p)$ to be infinite. Since $p_n \to p$, we can find an element $p_j \in \mathbf{S}_1$, now called y_2, such that $p_1 = y_1 \in \mathbf{I}^+(y_2)$. Repeating this argument always for the next element, we can find a subsequence $\mathbf{S}_2 = \{y_1, y_2, \cdots\}$ of $\mathbf{S}_1$ such that (a) holds:

(a) $y_i \in \mathbf{I}^+(y_{i+1})$ for all natural numbers i.

Using ordinary Minkowski coordinates with origin $p = 0 = (0,0,0,0)$, the standard Minkowski metric $g_{ik} = \delta_{ik} \cdot (-1)^{\delta_{i0}}$, and the corresponding Euclidian metric δ_{ik} for $i,k = 0,1,2,3$, we can draw a 3-sphere $\mathcal{S}^3$ centered at p with radius 1 with respect to δ_{ik}. If $\overrightarrow{py}_i$ is the ray from p through y_i, we consider the point set $\mathbf{T}_2 = \{t_i = \overrightarrow{py}_i \cap \mathcal{S}^3;\ y_i \in \mathbf{S}_2\}$ on the sphere. Since $\mathcal{S}^3$ is compact, $\mathbf{T}_2$ has an accumulation point q on $\mathcal{S}^3$ and there is a subsequence $\mathbf{S}_3$ of $\mathbf{S}_2$ such that $\mathbf{T}_3 = \{t_i = \overrightarrow{py}_i \cap \mathcal{S}^3;\ y_i \in \mathbf{S}_3\}$ converges to q. The ray $\overrightarrow{pq}$ is timelike or lightlike because of (2).

Now we select our final subsequence $\mathbf{S}_4 = \{z_1, z_2, \cdots\}$ of **S**: Let $|xy|$ be the distance between x and y with respect to δ_{ik}. We choose $z_1 = y_j$ such that $|t_j q| < e^{-1}$ and $|pz_1| < e^{-1}$. We assume $z_1, \ldots, z_{m-1}$ to be constructed such that

(b) $|(\overrightarrow{pz}_i \cap \mathcal{S}^3)q| < e^{-i}$,

(c) $|(\overrightarrow{z_i z_{i-1}} \cap \mathcal{S}^3)q| < e^{-i+1}$,

(d) $|pz_i| < e^{-i}$, for all $i < m$.

Since $\mathbf{S}_3 \to p$ and $\mathbf{T}_3 \to q$, we can find a point $z_m = y_j$ which satisfies (b)–(d) for $i = m$.

Now we draw the curve $\tilde{\tilde{c}}(\mathbf{S}_4)$ consisting of all pieces of straight lines $\overline{z_i z_{i+1}}$ including the limit point p. Because of (a) we can use the time coordinate as parameter along $\tilde{\tilde{c}}(\mathbf{S}_4)$. We obtain, that the path $\tilde{\tilde{c}}(t)$ with image $\tilde{\tilde{c}}(\mathbf{S}_4)$ is C^∞ at all points except possibly at p or $\mathbf{S}_4$. Because of (b) and (c) the path $\tilde{\tilde{c}}(t)$ is at least C^1 at $t = 0$. The "pathology" at $\mathbf{S}_4$ can be removed by the well-known C^∞-smoothing procedure; cf. Penrose,[4] p. 16. We perform the C^∞-smoothing at $z_n = \tilde{\tilde{c}}(t_n)$ such that the new path $c(t)$ coincides with $\tilde{\tilde{c}}(t)$ except at a sufficiently small neighborhood $\mathbf{V}_n$ of z_n. This can be done such that

(e) $c(t)$ is C^∞ and timelike in $\mathbf{V}_n$,

(f) $|(\overrightarrow{pc(t)} \cap \mathcal{S}^3)q| < e^{-n+1}$ for all points $c(t) \in \mathbf{V}_n$,

(g) $\tilde{\tilde{c}}(t_n) = c(t_n)$.

The resulting path $c(t)$ is timelike and future directed C^∞ at all points $c(t) \neq 0$ and C^1 timelike or lightlike and future directed at $p = 0$.

Proposition 2.1*: Under the assumptions of Proposition 2.1 and condition

(3*) if $\mathbf{S}' \subset \mathbf{S}$, the sequence of rays $\overrightarrow{pp}_n$ ($p_n \in \mathbf{S}'$) is not converging to a null line,

there is a C^1-world-line passing through p and infinitely many points of **S**. This world line is timelike everywhere.

Proof: Apply Proposition 2.1; Proposition 2.1* then follows directly from condition (3*).

Corollary 2.2: Let **M** be a space–time and g be a world line through p. Any sequence **S** of points on g with $\mathbf{S} \to p$ contains a subsequence **S'** of points p_n ($n \in \mathbb{N}$) such that

(1) $\mathbf{S}' \subset g$ and $p_n \to p$ with respect to the ordinary topology $\mathfrak{T}$,

(2) There is a C^1-path f on **M** passing through p and **S'**,

(3) This path is timelike everywhere except possibly at p where it is nonspacelike.

Proof: Take a convex normal neighborhood **U** of p and apply $\exp^{-1}$. Since g is a timelike C^0-path through p, the set $\exp^{-1}(\mathbf{S} \cap \mathbf{U})$ contains a sequence **S''** of points satisfying Proposition 2.1 for the Minkowski tangent space $\mathbf{T}_p$ at p. By Proposition 2.1 there is an (infinite) subsequence **S*** of **S''** on a C^1-world-line h in the tangent space $\mathbf{T}_p$. Since exp is a C^k-map (if $k \geq 1$ is the degree of differentiability of **M**), we obtain the required C^1-world-line $\exp \circ h$ if we restrict ourselves to a sufficiently small neighborhood of 0 in $\mathbf{T}_p$. This world line passes through p and $\mathbf{S}' = \exp(\mathbf{S}^*)$.

Corollary 2.2*: Let **M** be a space–time and g be a *strictly* timelike world line through p. Then the world line f constructed in Corollary 2.2 is timelike everywhere.

Proof: Apply Corollary 2.2 and Proposition 2.1*.

Proposition 2.3: Let **M** be the Minkowski space. For each $n > 1$ there is a sequence $\mathbf{S} = \mathbf{S}_n$ on **M** with the following properties:

(1) $\mathbf{S} \to p$,

(2) there is a C^{n-1}-path f passing through p and **S**, and f is strictly timelike,

(3) if g is a path which contains infinitely many points of **S**, then g is not C^n.

Proof: Using Minkowski coordinates we can restrict ourselves to the two-dimensional case of a $t-x$-Minkowski plane with Minkowski coordinates t (for time) and x (for space) at the origin $0=p$. If $n>1$, we choose the timelike path $f(t)=(t,t^{(2n-1)/2})$ for $0\leq t\leq\frac{1}{2}$ and select the point set $S=S_n=\{(1/k,(1/k)^{(2n-1)/2})=f(1/k);\ k\in\mathbb{N}\}$ of f and assume that there is a C^n-path $g=g(s)$ which passes through infinitely many points S' of S. Since f is timelike, we may assume g to be timelike (within a sufficiently small neighborhood of p). Therefore we can choose the time coordinate t as a C^n-parameter of g, i. e., $g=g(t)$. Next we calculate the mth derivative of $g(t)$ at $t=0$ for all $m\leq n$ under the assumption $g(t)\in C^n$. This can be done by selecting *one particular* sequence of differences of the mth order at 0 for $m=1,\ldots,n$. By induction we have $g^{(m-1)}(0)=0$ and therefore we obtain for the specially chosen sequence of the mth order,

$$\frac{t_k^{(2n-1)/2}}{t_k^m}=t_k^{n-m-1/2}=k^{m-n+1/2}\quad\text{for } [t_k=1/k,\ g(t_k)]\in S'.$$

Since $k\to\infty$, we get $g^{(m)}(0)=0$ if $m<n$ and $g^{(n)}(0)=\infty$, i. e., $g\notin C^n$.

In order to formulate our next proposition, we need certain (pairs of) subsets of a C^ω-manifold, which we call "*analytically exact*".

Let M be a C^ω-manifold. The pair $(\mathbf{X},p)$ will be called *analytically exact* if it satisfies the following conditions:

(1) $\mathbf{X}\subset M$, $\mathbf{X}\in\mathfrak{T}$ and $p\in M$,

(2) $p\in\overline{\mathbf{X}}=\mathfrak{T}$-closure of X in M,

(3) there is one and only one analytic curve $f\subset\overline{\mathbf{X}}$ containing p. This curve f will be called the (analytic) axis of $(\mathbf{X},p)$.

Of course there are infinitely many paths g with the same curve $f=g$.

Proposition 2.4: If $f:[0,1]\to\mathbf{M}$ is an analytic path of an analytic manifold **M** (of dimension four), there is a *standard* analytically exact pair $(\mathbf{X}_f,p)$ with axis f.

Proof: We choose local C^ω-coordinates at p. Therefore it is sufficient to prove the proposition in the case

(a) $M=\mathbb{R}^4$, and $p=0=(0,0,0,0)\in\mathbb{R}^4$.

We choose one of the coordinate axes to be the analytic path f. Therefore we may assume, without loss of generality,

(b) $f(x)=(x,0,0,0)$ for $0\leq x\leq 1$,

and we put

$$\mathbf{X}_f=\{(x_1,x_2,x_3,x_4);\ 0<x_1<1,$$
$$(x_2^2+x_3^2+x_4^2)^{1/2}<\exp(-1/x_1^2)\}.$$

Then the curve f is contained in $\overline{\mathbf{X}}_f$ and $\mathbf{X}_f\in\mathfrak{T}$. Therefore it is sufficient to check (3) of the "exactness" definition: Let $h:[0,1]\to\overline{\mathbf{X}}_f$ be analytical and $p\in h$. We calculate its kth derivative: Since $p\in\overline{\mathbf{X}}_f$, there is a sequence $s(n)\in\overline{\mathbf{X}}_f\cap h$ of points such that $s(n)\to p$. We have $s(n)=(x(n)_1,x(n)_2,x(n)_3,x(n)_4)$ with respect to the coordinate system. Since $s(n)\in h$, we have in case $k=1$,

$$\left|\left(\frac{\partial}{\partial x_j}h\right)(0)\right|=\left|\lim_{n\to\infty}\frac{x(n)_j}{x(n)_1}\right|\leq\lim_{n\to\infty}\left|\frac{x(n)_j}{x(n)_1}\right|$$

for $j=2,3,4$. Since $s(n)\in\overline{\mathbf{X}}_f$, we get $|x(n)_j|\leq\exp[-1/x(n)_1^2]$ and therefore

$$\left|\left(\frac{\partial}{\partial x_j}h\right)(0)\right|\leq\lim_{n\to\infty}\left|\frac{\exp[-1/x(n)_1^2]}{x(n)_1}\right|=0.$$

An analogous argument gives (for $k>1$)

$$\left|\frac{\partial^k}{\partial x_{j_1}\cdots\partial x_{j_k}}h(0)\right|\leq\lim_{n\to\infty}\left|\frac{\exp[-1/x(n)_1^2]}{p_k(x(n)_1)}\right|=0,$$

for some polynomial $p_k(x)$ and $j_i\neq 1$ for all i. Since h is analytic at $p=0$, the curve h coincides with the axis f.

3. RELATIONS BETWEEN ZEEMAN TOPOLOGIES ON A CURVED SPACE-TIME

The results of this section may be summarized as follows: The Zeeman topologies, introduced in Sec. 1 are related as shown in the diagram (+) of the Introduction. All inequalities are strict.

Theorem 3.1:

(a) $\mathfrak{P}_0=\mathfrak{P}_1\leq\mathfrak{P}_n\leq\mathfrak{P}_{n+1}\leq\mathfrak{P}_\infty\leq\mathfrak{P}_\omega\leq\mathfrak{P}_g$,

(a*) $\mathfrak{P}_0^*=\mathfrak{P}_1^*\leq\mathfrak{P}_n^*\leq\mathfrak{P}_{n+1}^*\leq\mathfrak{P}_\infty^*\leq\mathfrak{P}_\omega^*\leq\mathfrak{P}_g^*$,

(b) $\mathfrak{Z}_0=\mathfrak{Z}_1\leq\mathfrak{Z}_n\leq\mathfrak{Z}_{n+1}\leq\mathfrak{Z}_\infty\leq\mathfrak{Z}_\omega\leq\mathfrak{Z}_g$,

(b*) $\mathfrak{Z}_0^*=\mathfrak{Z}_1^*\leq\mathfrak{Z}_n^*\leq\mathfrak{Z}_{n+1}^*\leq\mathfrak{Z}_\infty^*\leq\mathfrak{Z}_\omega^*\leq\mathfrak{Z}_g^*$

for all natural numbers n,

(c) $\mathfrak{P}_n\leq\mathfrak{P}_n^*$, $\mathfrak{P}_g=\mathfrak{P}_g^*$,

(c*) $\mathfrak{Z}_n\leq\mathfrak{Z}_n^*$, $\mathfrak{Z}_g=\mathfrak{Z}_g^*$, $\mathfrak{Z}_n\leq\mathfrak{P}_n$, for $n=1,2,\ldots,\infty,\omega$.

Proof: By definition we get

$$\mathfrak{P}_0\leq\mathfrak{P}_1\leq\cdots\leq\mathfrak{P}_n\leq\mathfrak{P}_{n+1}\cdots\leq\mathfrak{P}_\infty\leq\mathfrak{P}_\omega.$$

Let be $\mathbf{X}\in\mathfrak{P}_1$ and let us assume $\mathbf{X}\notin\mathfrak{P}_0$. There is a timelike C^0-path g such that $g\cap\mathbf{X}$ is not open in g. Therefore, $g\setminus(g\cap\mathbf{X})$ is not closed, i. e., there is a sequence $S=\{p_1,p_2,\cdots\}$ of points on $g\setminus(g\cap\mathbf{X})$ which converges to a point $p\in g\cap\mathbf{X}$. Because of Corollary 2.2 there is a subsequence S' of S which converges to p and is contained in a C^1-curve h which is timelike at all points $x\neq p$ and timelike or possibly lightlike at p. Since $\mathbf{X}\in\mathfrak{P}_1$ we get $\mathbf{X}\cap h=\mathbf{T}\cap h$ for some $\mathbf{T}\in\mathfrak{T}$. Since $S'\subset S\cap g$ and $S\cap\mathbf{X}=\emptyset$ by construction, we get $S'\cap\mathbf{X}=\emptyset$. If we assume $x\in S'\cap\mathbf{T}\neq\emptyset$, we get $x\in g$ (since $x\in S'$) and $x\in h\cap\mathbf{T}=h\cap\mathbf{X}$, i. e., $x\in\mathbf{X}\cap S'\neq\emptyset$ which is a contradiction. Therefore $S'\subset M\setminus\mathbf{T}$ which is $\mathfrak{T}$ closed. Since $S'\to p$ with respect to $\mathfrak{T}$, it follows that $p\in M\setminus\mathbf{T}$, i. e., $p\notin\mathbf{T}$ or $p\notin\mathbf{T}\cap h=\mathbf{X}\cap h$. Since $p\in h$, we get $p\notin\mathbf{X}$, which contradicts $p\in g\cap\mathbf{X}\subset\mathbf{X}$. Therefore, $\mathbf{X}\in\mathfrak{P}_0$, and (a) is shown.

(a*) follows as (a): Since the C^0-curve g (as above) is strictly timelike at p, the corresponding sequence S' satisfies assumption (3*) of Proposition 2.1*. Therefore, h is timelike at p, and we can apply the arguments in (a).

(b) and (b*) follow as in (a) and (a*) by always adding the "hypersurface-condition (+++)."

Since the sets of test curves are ordered by inclusion in oposite directions, the ordering of (c) holds automatically. Since geodesics are either null or timelike or spacelike, the sets of test curves for $\mathfrak{P}_g$ and $\mathfrak{P}_g^*$ coincide. Therefore $\mathfrak{P}_g = \mathfrak{P}_g^*$ and (c) are shown.

(c*) follows as (c).

All inclusions of Theorem 3.1 are strict, i.e.,

Lemma 3.2:

(a) $\mathfrak{P}_n \neq \mathfrak{P}_{n+1}$, $\mathfrak{P}_\infty \neq \mathfrak{P}_\omega \neq \mathfrak{P}_g$,

(a*) $\mathfrak{P}_n^* \neq \mathfrak{P}_{n+1}^*$, $\mathfrak{P}_\infty^* \neq \mathfrak{P}_\omega^* \neq \mathfrak{P}_g$,

(b) $\mathfrak{Z}_n \neq \mathfrak{Z}_{n+1}$, $\mathfrak{Z}_\infty \neq \mathfrak{Z}_\omega \neq \mathfrak{Z}_g$,

(b*) $\mathfrak{Z}_n^* \neq \mathfrak{Z}_{n+1}^*$, $\mathfrak{Z}_\infty^* \neq \mathfrak{Z}_\omega^* \neq \mathfrak{Z}_g$,

for all natural numbers n,

(c) $\mathfrak{P}_n \neq \mathfrak{P}_n^*$,

(c*) $\mathfrak{P}_n \neq \mathfrak{P}_n^*$, $\mathfrak{Z}_n \neq \mathfrak{P}_n$, for $n = 1, 2, \ldots, \infty, \omega$.

Proof: Let U be a simple convex (normal) neighborhood at $p \in \mathbf{M}$ and *exp* the exponential map of a region V of the tangent space $\mathbf{T}_p$ onto U. (a), (a*), (b), (b*):

Case $n \neq \infty, \neq \omega$: In $\mathbf{T}_p$ we can find a sequence $\mathbf{S} = \mathbf{S}_{n+1}$ of points satisfying Proposition 2.3 for $p = 0$. Therefore, $\mathbf{S}^* = exp(\mathbf{S} \cap \mathbf{V})$ satisfies the same conditions within the manifold (if M is sufficiently smooth, as always assumed; cf. Sec. 1). Therefore, $|\mathbf{S}^* \cap g| < \infty$ for all timelike C^{n+1} paths g of M. Hence $\mathbf{U}^* = \boldsymbol{U} \setminus \mathbf{S}^* \in \mathfrak{P}_{n+1}^*$ and, since f is chosen to be strictly timelike in Proposition 2.3, $\mathbf{U}^* \in \mathfrak{Z}_{n+1}$ ($< \mathfrak{Z}_{n+1}^*$, $< \mathfrak{P}_{n+1}$, $< \mathfrak{P}_{n+1}^*$). However, $\mathbf{U}^* \notin \mathfrak{P}_n^*$ (and therefore not in $\mathfrak{Z}_n^*, \mathfrak{Z}_n, \mathfrak{P}_n$) since there is a strictly timelike C^n-path f with infinite $\mathbf{S}^* \cap f$ which follows from Proposition 2.3(2).

Case $\mathfrak{P}_\infty \neq \mathfrak{P}_\omega$: Choose a strictly timelike analytic path f through p. There is a standard analytically exact pair $(\mathbf{X}_f, p)$; cf. Proposition 2.4. Choose a sequence S of points in $\mathbf{X}_f$ on a C^∞-path g (which can be arranged easily) converging to p such that $|f \cap \mathbf{S}| < \infty$. If c is any analytic curve with $|c \cap \mathbf{S}| = \infty$ then $p \in c$ and $c = f$ by the definition of "analytical exactness". Therefore $\infty = |c \cap \mathbf{S}| = |f \cap \mathbf{S}| < \infty$ is a contradiction. We get from the definition of Zeeman topology: $\mathbf{U}^* = \boldsymbol{U} \setminus \mathbf{S} \in \mathfrak{Z}_\omega$ ($< \mathfrak{P}_\omega$, $< \mathfrak{P}_\omega^*$, $< \mathfrak{Z}_\omega^*$). Since $\mathbf{S} \subset g \in C^\infty$, we obtain on the other hand $\mathbf{U}^* \notin \mathfrak{P}_\infty^*$ ($> \mathfrak{P}_\infty$, $> \mathfrak{Z}_\infty$, $> \mathfrak{Z}_\infty^*$).

Case $\mathfrak{P}_\omega \neq \mathfrak{P}_g$: Choose an analytic path f which is strictly timelike at p and which is not a piece of a geodesic at p. Then it is possible to find a sequence S of points on f such that each geodesic contains only finitely many of them; cf. R. Göbel.[2] Hence the set $\mathbf{U}^* = \boldsymbol{U} \setminus \mathbf{S}$ belongs to $\mathfrak{P}_g \cap \mathfrak{Z}_g$ but not to $\mathfrak{P}_\omega$ or $\mathfrak{Z}_\omega$, or $\mathfrak{Z}_\omega^*$ or $\mathfrak{P}_\omega^*$.

Case (c) *and* (c*): Take any C^n-world-line which is lightlike at p and timelike at all other points. Choose a sequence S on this line which converges to p. Then $\mathbf{U}^* = \boldsymbol{U} \setminus \mathbf{S} \in \mathfrak{Z}_n^*$ ($\in \mathfrak{P}_n^*$) but $\mathbf{U}^* \notin \mathfrak{P}_n$ ($\notin \mathfrak{Z}_n$) for all $n = 1, 2, \ldots, \infty, \omega$.

The next proposition is an analog of Theorem 1 in HKM.[3] It has the *advantage of being "trivial" and leading to the main results* (in particular to Theorem 4) in HKM[3]; it can be used as a substitute of Theorem 1 in that paper. It follows immediately from Proposition 3.3 that $\mathfrak{P}$-topologies, restricted to the light cones are discrete. From the topological point of view, the structure of the Zeeman topologies does *not* become clearer under Proposition 3.3 (in spite of HKM Theorem 1) which is its disadvantage.

Let $\mathbf{K}(p, \epsilon)$ be defined as in HKM,[3] i.e., $\mathbf{K}(p, \epsilon)$ consists of the interior of the light cone at p within an "ϵ-neighborhood" of p, and again we include the point p. Then let $\mathfrak{B}_k$ consist of all $\mathfrak{P}_k$-open sets $\mathbf{K}(p, \epsilon, \mathbf{S})$ contained in arbitrary sets $\mathbf{K}(p, \epsilon)$. The "index" S in $\mathbf{K}(p, \epsilon, \mathbf{S})$ can be thought of as a $\mathfrak{P}_k$-closed subset of $\mathbf{K}(p, \epsilon)$.

Proposition 3.3: The set $\mathfrak{B}_k$ is a basis for the topology $\mathfrak{P}_k$.

Proof: If $\mathbf{X} \in \mathfrak{B}_k$, then $\mathbf{X} \in \mathfrak{P}_k$ by definition. Therefore let $p \in \mathbf{X} \in \mathfrak{P}_k$. Since $p \in \mathbf{K}(p, \epsilon) \in \mathfrak{P}_k$, we get $p \in \mathbf{K}(p, \epsilon) \cap \mathbf{X} = \mathbf{B} \in \mathfrak{P}_k$ and therefore $\mathbf{B} \in \mathfrak{B}_k$. Hence $\mathfrak{B}_k$ generates $\mathfrak{P}_k$.

From the topological point of view we would like to remark (without proofs) that:

(a) all topologies in consideration are Hausdorff, locally and globally path connected (cf. HKM,[3] Sec. 4);

(b) $\mathfrak{P}_k$ is countable if and only if $k = 0$ and $k = 1$; cf. HKM[3];

(c) none of the topologies is regular or normal or locally compact or paracompact (since $\mathfrak{P}_0$ is not; cf. HKM,[3] Theorem 3, and $\mathfrak{Z}_0$ is not either).

4. WORLD LINES WITH RESPECT TO ZEEMAN TOPOLOGIES

First we will consider *world lines* (cf. Sec. 1) with respect to the different Zeeman topologies. Without loss of generality we assume world lines to be 1–1 maps from $[0, 1]$ into the space–time M which are order preserving, cf. Sec. 1. This could easily be generalized to *locally* 1–1 maps from open connected subsets of $\mathbb{R}$ into M; cf. HKM.[3] We will restrict ourselves (cf. Sec. 1) to the following groups of topologies only:

(a) $\mathfrak{P}_0 = \mathfrak{P}_1$, $\mathfrak{P}_0^* = \mathfrak{P}_1^*$, $\mathfrak{Z}_0 = \mathfrak{Z}_1$, $\mathfrak{Z}_0^* = \mathfrak{Z}_1^*$,

(b) $\mathfrak{P}_\omega$, $\mathfrak{P}_\omega^*$, $\mathfrak{Z}_\omega$, $\mathfrak{Z}_\omega^*$,

(c) $\mathfrak{P}_g$, $\mathfrak{Z}_g$.

The last case (c) was solved by Göbel[2] (Corollary 3.4): *World lines are broken timelike future directed geodesics.* The case $\mathfrak{P}_0$ has been investigated in HKM.[3] Using the same techniques (apply Proposition 3.3 and Theorem 2 or Proposition 4.1 in HKM[3]) we derive the *same* result for $\mathfrak{Z}_0$ as for $\mathfrak{P}_0$.

Theorem 4.1: For a 1–1 map $f: [0, 1] \to \mathbf{M}$ are equivalent:

(1) f is a world line with respect to $\mathfrak{P}_0$ or $\mathfrak{Z}_0$,

(2) f is a world line with respect to $\mathfrak{T}$ (not necessarily strictly timelike).

There is no difference between $\mathfrak{P}_0$, $\mathfrak{Z}_0$, and $\mathfrak{T}$ for world lines! Next we consider $\mathfrak{P}_0^*$ and $\mathfrak{Z}_0^*$.

Theorem 4.2: For a 1–1 map $f:[0,1]\to \mathbf{M}$ are equivalent:

(1) f is a world line with respect to $\mathfrak{P}_0^*$ or $\mathfrak{Z}_0^*$,

(2) f is a world line with respect to $\mathfrak{T}$ but in addition strictly timelike.

Therefore world lines continuously reaching the light velocity are singled out.

Proof: (2) → (1): $\mathfrak{P}_0^*$ [$\mathfrak{Z}_0^*$]-continuity of world lines f which are strictly timelike everywhere follows directly from the definition of $\mathfrak{P}_0^*$ [$\mathfrak{Z}_0^*$].

(1) → (2): Since $\mathfrak{P}_0 < \mathfrak{Z}_0^*$ (Theorem 3.1) we may apply Theorem 4.1 and f is a world line with respect to $\mathfrak{T}$. If f were not strictly timelike at $p=f(t_0)\in f$, we could find a sequence S of points $p_n \neq p$ within a simple neighborhood $\boldsymbol{V}$ of p such that the geodesics $\overline{pp_n}$ with $p_n \in \mathbf{S}$ are approaching the light cone. Therefore $\mathbf{V}^* = \boldsymbol{V}\setminus \mathbf{S} \in \mathfrak{P}_0^*$ but $\mathbf{V}^* \notin \mathfrak{P}_0$ [cf. Proof of Lemma 3.2(c)]. In addition $\mathbf{V}^* \in \mathfrak{Z}_0^* \setminus \mathfrak{Z}_0$. Since $p\in \mathbf{V}^*$ and $f^{-1}(p_n)\to t_0$ we derive the contradiction to the $\mathfrak{P}_0^*$-continuity (or the $\mathfrak{Z}_0^*$-continuity) of f at p.

Next we consider the cases $\mathfrak{P}_\omega/\mathfrak{Z}_\omega$. The restriction to analytic manifolds is *no real restriction* as it follows from the famous embedding theorem by Hassler and Whitney, cf. Hawking and Ellis,[9] p. 58! We derive the somewhat surprising

Theorem 4.3: For a 1–1 map $f:[0,1]\to \mathbf{M}$ are equivalent:

(1) f is a world line with respect to $\mathfrak{P}_\omega$ or $\mathfrak{Z}_\omega$,

(2) f is a broken analytic curve, which is timelike almost everywhere (except for possibly finitely many null tangents) and future directed.

Proof: (2) → (1) follows immediately from the definitions. (1) → (2): Let be $p=f(0)\in f$ and assume f not to be analytic at p. There are two possibilities which may occur:

(i) There is no analytic curve g which contains a sequence S of points $\neq p$ such that $\mathbf{S}\subset g\cap f$ and $\mathbf{S}\to p$.

(ii) There is an analytic curve g such that $g\cap f\setminus\{p\}$ contains a sequence $\mathbf{S}\to p$.

In the first case we choose an arbitrary sequence $\mathbf{S}\subset f$ such that $\mathbf{S}\to p$. It follows from (i) that $|\mathbf{S}\cap g|<\infty$ for all analytic curves g. Therefore $\mathbf{U}^* = \boldsymbol{U}\setminus \mathbf{S} \in \mathfrak{P}_\omega$ [$\in \mathfrak{Z}_\omega$] if $\mathbf{U}$ is a simple neighborhood of p. Since $p\in \mathbf{U}^*$ and $\mathbf{S}\to p$, this contradicts $\mathfrak{P}_\omega$-continuity (or $\mathfrak{Z}_\omega$-continuity) of f at p.

In the second case we choose the standard exact analytic pair $(\mathbf{X}_g, p)$ with axis g; cf. Proposition 2.4. We have $\mathbf{S}\subset g\cap f$ and $\mathbf{S}\to p$ from (ii) and $\mathbf{X}_g \in \mathfrak{T}$ and f is connected. Therefore f must have points in common with $\mathbf{X}_g \setminus g$ arbitrarily close to p since g is analytic at p but f is not. We obtain a sequence $\mathbf{S}^* \subset f\cap \mathbf{X}_g \setminus g$ and $\mathbf{S}^*\to p$. From condition (3) of "exactness" we know that $|\mathbf{S}^*\cap h|<\infty$ for all analytic curves h. Therefore, $\mathbf{U}^* = \boldsymbol{U}\setminus \mathbf{S}^*$ is a $\mathfrak{P}_\omega$-neighborhood (and a $\mathfrak{Z}_\omega$-neighborhood) of p, contradicting $\mathfrak{P}_\omega$-continuity ($\mathfrak{Z}_\omega$-continuity) of f at p. Therefore case (ii) cannot occur either and f is piecewise analytic at p. This argument can be repeated for any point p on f. Using the fact that f is a compact set, (2) follows immediately; cf. Göbel[2] (for the explicit argument).

Corollary 4.4: For a 1–1 map $f:[0,1]\to \mathbf{M}$ are equivalent:

(1) f is a world line with respect to $\mathfrak{P}_\omega^*$ or $\mathfrak{Z}_\omega^*$,

(2) f is a broken analytic curve which is timelike and future directed everywhere.

Proof: (2) → (1) follows from the definitions. (1) → (2): Since $\mathfrak{P}_\omega < \mathfrak{P}_\omega^*$ we obtain from Theorem 4.3 that f is piecewise analytic. That null tangents do not occur follows from Theorem 4.2 since $\mathfrak{P}_0^* < \mathfrak{P}_\omega^*$. The same argument holds for $\mathfrak{Z}_\omega^*$.

5. FEYNMAN TRACKS FOR "$\mathfrak{P}$-TYPE" TOPOLOGIES

First we remark, that there are "arbitrarily wild" $\mathfrak{Z}$-continuous maps $f:[0,1]\to \mathbf{M}$ if we drop the order preserving for *all* "$\mathfrak{Z}$-type" topologies. Therefore, in the following we restrict ourselves to the $\mathfrak{P}$-type. We shall characterize all $\mathfrak{P}$-continuous maps $f:[0,1]\to \mathbf{M}$ with respect to $\mathfrak{P}_0^*$, $\mathfrak{P}_\omega$, $\mathfrak{P}_\omega^*$, $\mathfrak{P}_g$. This has been done for $\mathfrak{P}_0$ in HKM[3] (Theorem 2) and in order to avoid overlappings, we will make use of this result continuously.

We call a curve f an *analytic Feynman track* if f is connected and consists of finitely many pieces of future directed or past directed analytic world-lines. The track will be called *strict* if no null tangents occur. We mention without proof, that $\mathfrak{P}_0^*$-continuous curves are Feynman tracks in the (more general) sense of HKM,[3] which are in addition strict; cf. HKM.[3]

Theorem 5.1: Let $\mathbf{M}$ be an analytic space–time and $f:[0,1]\to \mathbf{M}$ a 1–1 map. Then f is $\mathfrak{P}_\omega$-continuous if and only if f is an analytic Feynman track.

Proof: "⇐" follows by the definitions. "⇒": Since $\mathfrak{P}_0 < \mathfrak{P}_\omega$, we already know from Theorem 2 in HKM[3] that f is a Feynman track in the sense of HKM.[3] Therefore four possibilities arise: Either f is order preserving or reversing in a neighborhood $\mathbf{U}$ of $t_0\in[0,1]$ or $f(\mathbf{U})\subset \mathbf{I}^+(p)\cup\{p\}$, or $f(\mathbf{U})\subset \mathbf{I}^-(p)\cup\{p\}$. In the first two cases we apply Theorem 4.3 to obtain that $f(\mathbf{U})$ is piecewise analytic at $p=f(t_0)$. In the third case (and similarly in the fourth case) we restrict ourselves to $\mathbf{U}^+ = \{t\in \mathbf{U}, t\geq t_0\}$ and obtain that f is order preserving *at* t_0, i.e., $p\ll f(t)$ for all $t\in \mathbf{U}^+$ sufficiently close to t_0. Using the argument in the proof "(1) → (2)" of Theorem 4.3, we derive that there is a $t'\in \boldsymbol{U}^+\setminus t_0$ such that $f[t_0,t']$ is an analytic curve. After this is shown for all points $p\in f$ we apply that $f=f[0,1]$ is compact. Therefore f is a (broken) analytic Feynman track.

If we combine Theorem 5.1, Corollary 4.4, and Theorem 3.1(c) (for $n=\omega$), we derive Corollary 5.2(a).

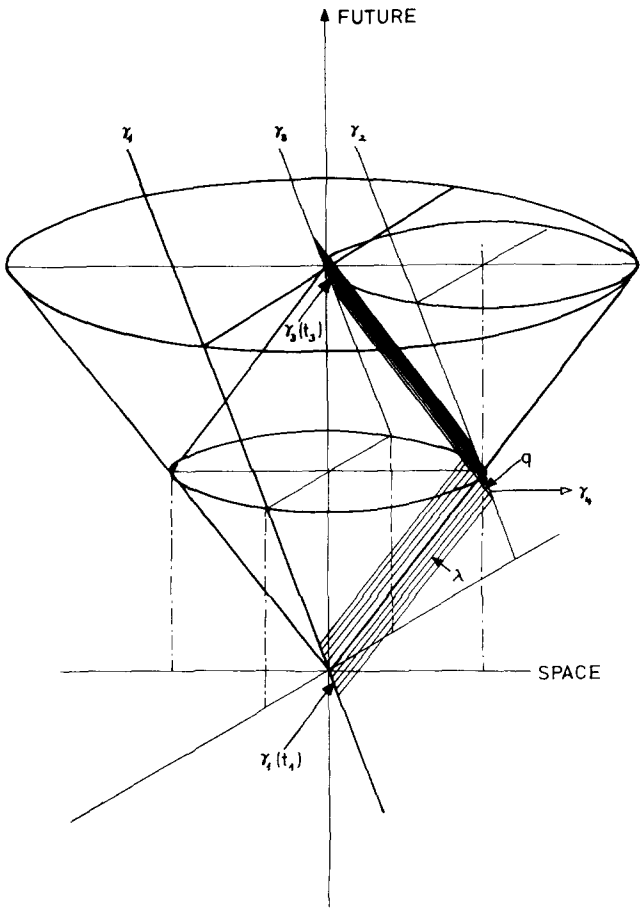

FIG. 1.

Using Theorem 5.1, Theorem 3.1(a), and Corollary 3.4 in Göbel,[2] we obtain Corollary 5.2(b).

Corollary 5.2: Let $f[0,1] \to \mathbf{M}$ be a 1—1 map into the space—time $\mathbf{M}$.

(a) If $\mathbf{M}$ is analytic, then f is $\mathfrak{P}^*_\omega$-continuous if and only if f is a strictly timelike analytic Feynman track.

(b) f is $\mathfrak{P}_g$-continuous if and only if f is a piecewise future and past directed geodesic Feynman track which is timelike everywhere.

6. THE GROUP OF HOMEOMORPHISMS

From now on we will assume that the Lorentz metric on our space—time $\mathbf{M}$ is C^∞ or C^ω if $\mathbf{M}$ is C^∞ or C^ω respectively. The following proposition will be used to prove Theorem 6.2.

Proposition 6.1: A $\mathfrak{T}$-homeomorphism $h: \mathbf{M} \to \mathbf{M}'$ of an analytic space—time $(\mathbf{M}, g)$ onto an analytic space—time $(\mathbf{M}', g')$, which takes null geodesic curves to null geodesic curves (as point sets), is an analytic diffeomorphism.

Remark: $(\mathbf{M}, g)$, $(\mathbf{M}', g')$ are analytic manifolds with analytic metric fields g and g' respectively. Proposition 6.1 is the "*analytic analog*" of *Hawking's* Theorem 5 in HKM.[3] We apply the method given there and refer for notations to HKM[3] (Proof of Theorem 5) and Fig. 1 which illustrates the idea of the construction. The analytic diffeomorphism is in addition conformal. The definition of *conformal maps* used below—known in mathematics at least since 1847 (Liouville[10])—may be found in Hawking and Ellis[9] (p. 42) or Göbel[2] (Sec. 2). Conformal maps are (differentiable and) angle preserving within the tangent spaces of the space—time.

Proof: We already know from HKM[3] (proof of Theorem 5) that the maps $\hat{h}_i: \mathbf{F}_i \to \tilde{\mathbf{F}}_i$ are C^∞ for $i = 1, 2, 3, 4$. Parametrizing points q on γ_2 by parameters t_1 and t_3 of γ_1 and γ_3, the functional equation $\hat{h}_4(\psi_2(t_1, t_3)) = \tilde{\psi}_2(\hat{h}_1(t_1), \hat{h}_3(t_3))$ is derived in HKM[3] (we put $\psi = \psi_2$), cf. Fig. 1. Permuting the suffix (and parametrizing successively γ_i for $i = 1, 2, 3, 4$) we obtain four functional equations, which have been used implicitly in HKM[3]:

$$\begin{aligned}
&\hat{h}_1(\psi_3(t_2, t_4)) = \tilde{\psi}_3(\hat{h}_2(t_2), \hat{h}_1(t_4)),\\
&\hat{h}_2(\psi_4(t_3, t_1)) = \tilde{\psi}_4(\hat{h}_3(t_3), \hat{h}_1(t_1)),\\
&\hat{h}_3(\psi_1(t_4, t_2)) = \tilde{\psi}_1(\hat{h}_4(t_4), \hat{h}_2(t_2)),\\
&\hat{h}_4(\psi_2(t_1, t_3)) = \tilde{\psi}_2(\hat{h}_1(t_1), \hat{h}_3(t_3)).
\end{aligned} \tag{1}$$

The functions ψ and $\tilde{\psi}$ constructed with "small" pieces of null geodesics are analytic with respect to a sufficiently small parameter domain as follows from an elementary result of differential (equations) geometry; cf. Hicks[11] (p. 59). Hence (1) represents an analytic system of equations for the unknown functions $\hat{h}_1, \hat{h}_2, \hat{h}_3, \hat{h}_4$ which may be written in the form

$$\begin{aligned}
&\text{(a)}\quad F_1 = \hat{h}_1(\psi_3(t_2, t_4)) - \tilde{\psi}_3(\hat{h}_2(t_2), \hat{h}_4(t_4)) = 0,\\
&\text{(b)}\quad F_2 = \hat{h}_2(\psi_4(t_3, t_1)) - \tilde{\psi}_4(\hat{h}_3(t_3), \hat{h}_1(t_1)) = 0,\\
&\text{(c)}\quad F_3 = \hat{h}_3(\psi_1(t_4, t_2)) - \tilde{\psi}_1(\hat{h}_4(t_4), \hat{h}_2(t_2)) = 0,\\
&\text{(d)}\quad F_4 = \hat{h}_4(\psi_2(t_1, t_3)) - \tilde{\psi}_2(\hat{h}_1(t_1), \hat{h}_3(t_3)) = 0.
\end{aligned} \tag{2}$$

Comparison with Minkowski space within a sufficiently small region shows that

$$\text{(i)}\quad \frac{\partial\tilde{\psi}_1(\tilde{t}_4, \tilde{t}_2)}{\partial\tilde{t}_2} \neq 0, \quad \frac{\partial\tilde{\psi}_2(\tilde{t}_1, \tilde{t}_3)}{\partial\tilde{t}_3} \neq 0,$$
$$\frac{\partial\tilde{\psi}_3(\tilde{t}_2, \tilde{t}_4)}{\partial\tilde{t}_4} \neq 0, \quad \frac{\partial\tilde{\psi}_4(\tilde{t}_3, \tilde{t}_1)}{\partial\tilde{t}_1} \neq 0,$$

and

$$\text{(ii)}\quad \frac{\partial\tilde{\psi}_1}{\partial\tilde{t}_4}\frac{\partial\tilde{t}_4}{\partial\tilde{t}_2} \neq \frac{\partial\tilde{\psi}_1}{\partial\tilde{t}_2}, \quad \frac{\partial\tilde{\psi}_2}{\partial\tilde{t}_3}\frac{\partial\tilde{t}_3}{\partial\tilde{t}_1} \neq \frac{\partial\tilde{\psi}_2}{\partial\tilde{t}_1}.$$

The conditions (i) have been used in HKM[3] implicitly and the last one is given explicitly, cf. HKM[3] [Proof of Theorem 5, (4)]. With the third condition of (i) we calculate from (2a)

$$\frac{\partial F_1}{\partial\hat{h}_4(t_4)} = -\frac{\partial\tilde{\psi}_3}{\partial\hat{h}_4(t_4)} \neq 0,$$

and by construction of ψ_3 and $\tilde{\psi}_3$ there are some parameters t_2^0, t_4^0 and t_3^0 $[= \psi_3(t_2^0, t_4^0)]$ such that $\hat{h}_1(t_3^0) - \tilde{\psi}_3(\hat{h}_2(t_2^0), \hat{h}_4(t_4^0)) = 0$; cf. Fig. 1. Hence (2a) can be solved, i.e., $\hat{h}_4(t_4) = f_1(\hat{h}_1(t_1), \hat{h}_3(t_3))$ is an analytic function by the implicit function theorem for a sufficiently small domain; cf. Narasimhan,[12] p. 15, 17, 18. In fact

$$\hat{h}'(t_4) = \left(\frac{\partial\tilde{\psi}_3}{\partial\hat{h}_4(t_4)}\right)^{-1} \hat{h}_1'(\psi_3(t_2, t_4))\frac{\partial\psi_3}{\partial t_4};$$

cf. HKM[3] (II). Substituting into (2b), (2c), and (2d) we obtain three analytic equations for the remaining functions $\hat{h}_2, \hat{h}_3, \hat{h}_4$. Repeating this method two times, (2d) can be reduced with (i) and (ii) to an analytic equation which determines $\hat{h}_3(t_3)$. Hence $\hat{h}_3(t_3)$ depends analytically on t_3. By symmetry we derive that $\hat{h}_i(t_i)$ is an analytic function for each $i=1,2,3,4$. Hence h is analytic too, repeating the final arguments in HKM[3] (Proof of Theorem 5).

Remark: If $h: M \to M'$ is conformal and C^2 and (M,g), (M',g') are analytic, we can derive the analytic differential equation, describing h locally:

$$(+)\quad g_{ik}(y^0,\ldots,y^3)\frac{\partial x^i(y^0,\ldots,y^3)}{\partial y^j}\,\frac{\partial x^k(y^0,\ldots,y^3)}{\partial y^s} = \Omega^2(y^0,\ldots,y^3)g_{js}(y^0,\ldots,y^3).$$

Such an equation has only analytic solutions (since $\dim M \geq 3$), as follows from Proposition 6.1. However it seems to be difficult to derive such a result directly from (+), which shows the insufficient knowledge on partial differential equations. That (+) has only analytic solutions in *flat space*, follows after lengthy calculations in the 100 year old paper by Beez[13] (and follows of course in the end from Liouville's theorem too).

Theorem 6.2: Let M_1 and M_2 be two strongly causal space–times which are C^∞ or $[C^\omega]$. Let $\mathfrak{X}_1$ and $\mathfrak{X}_2$ be, respectively, *one* of the topologies $\mathfrak{P}_i$ or $\mathfrak{P}_i^*$ $(i=0,1,\ldots,\infty,\omega)$ respectively on M_1 and on M_2. The following conditions are equivalent:

(1) $h: (M_1,\mathfrak{X}_1) \to (M_2,\mathfrak{X}_2)$ is an $\mathfrak{X}$-homeomorphism,

(2) $h: M_1 \to M_2$ is a conformal map which is C^∞ [or C^ω].

Remark: The definition of "strongly causal" can be found in Hawking and Ellis[19] (p. 192) or in Penrose[4] (p. 34, Theorem 4.24). Loosely speaking, space–times with "almost closed" world lines are excluded.

Proof: (2) → (1): Timelike C^k-curves are mapped onto timelike C^k-curves under a conformal C^∞- [or C^ω-] map for each degree of differentiability k. Using the definition of $\mathfrak{X}_1$ and $\mathfrak{X}_2$, we obtain that $\mathfrak{X}_1$-open sets are mapped onto $\mathfrak{X}_2$-open sets. Hence conformal maps are $\mathfrak{X}$-homeomorphisms.

(1) → (2): Let f be a timelike $\mathfrak{X}_1$-continuous world line of M_1. Since $\mathfrak{P}_0 \leq \mathfrak{X}_1$, it follows that f is timelike and $\mathfrak{P}_0$-continuous. Using HKM[3] (Theorem 4 and Proposition 5.1), we get that $h \circ f$ is timelike and $\mathfrak{P}_0$-continuous. Since h is an $\mathfrak{X}$-homeomorphism, $h \circ f$ is in addition to that $\mathfrak{X}_2$-continuous. Therefore $\mathfrak{X}_1$-continuous timelike curves are mapped onto $\mathfrak{X}_2$-continuous timelike curves. Therefore h is a causal map in the sense of Göbel,[2,14] i.e.,

(C) $\{x \ll y$ or $y \ll x\}$ if and only if $\{h(x) \ll h(y)$ or $h(y) \ll h(x)\}$.

It has been shown in Göbel[14] [cf. also Göbel[2] (Lemma 5.4) or HKM[3] (Proposition 5.4)] that causal maps are orthochronal or antiorthochronal, i.e.,

(O) $x \ll y$ if and only if $h(x) \ll h(y)$, (*orthochronal*),

(A) $x \ll y$ if and only if $h(y) \ll h(x)$, (*antiorthochronal*).

Therefore, h maps lightlike geodesics onto lightlike geodesics as shown in Göbel[2] (Sec. 5, p. 305) and HKM[3] (Proposition 6.1). Hence h is a conformal C^∞-map as follows from a theorem by Hawking[15] which is now published in HKM[3] (Theorem 5). In the case $n=\omega$ we apply Proposition 6.1 in order to get h to be analytic, too.

Corollary 6.3: Let M_1 and M_2 be two strongly causal space–times endowed with the topologies $\mathfrak{X}_1$ and $\mathfrak{X}_2$ both of the type $\mathfrak{P}_g$. Then the following conditions are equivalent:

(1) $h: (M_1,\mathfrak{X}_1) \to (M_2,\mathfrak{X}_2)$ is a $\mathfrak{P}_g$-homeomorphism.

(2) $h: M_1 \to M_2$ is a homothetic map, i.e., a conformal map with a *constant* conformal factor.

Remark: Corollary 6.3 has been conjectured by Zeeman in Ref. 1, p. 169, for $M_1 = M_2 =$ Minkowski space, which is a special strongly causal space–time. This proof should be substituted for the attempt by Nanda[7] in this journal; cf. Geroch.[8]

Proof: (1) → (2): Since $\mathfrak{P}_0 < \mathfrak{P}_g$, we obtain from Theorem 6.2 that h is conformal. From Corollary 5.2(b) we know that timelike geodesics are mapped onto broken timelike geodesics under h. Since h is conformal, it is differentiable, hence "image-geodesics" are no longer broken. Therefore timelike geodesics are mapped onto timelike geodesics under h. By application of Göbel[2] (Proposition 5.8) we derive that h is a homothetic map.

(2) → (1): Since timelike geodesics are mapped onto timelike geodesics under a homothetic transformation h [cf. Göbel[2] (Proposition 5.8)], the condition (1) follows immediately from the definition of $\mathfrak{P}_g$.

In the case $\mathfrak{Z}_\omega$ we can drop the assumption "strongly causal" using Göbel[2] (Corollary 5.7, p. 302). Using the techniques in Ref. 2 and in Sec. 4 of this paper we can derive the following theorem.

Theorem 6.4: Let M_1, M_2 be two analytic space–times endowed with the topologies $\mathfrak{X}_1$ and $\mathfrak{X}_2$ of the type $\mathfrak{Z}_\omega$. The following conditions are equivalent:

(1) $h: (M_1,\mathfrak{X}_1) \to (M_2,\mathfrak{X}_2)$ is a $\mathfrak{Z}_\omega$-homeomorphism.

(2) $h: M_1 \to M_2$ is an (analytic) conformal map between M_1 and M_2.

In the case of Minkowski space Theorem 6.4, Part (2) can be specialized even more, which is due to Liouville[10] and Lie[16]: The conformal maps of Theorem 6.4, Part (2) are "classical conformal" maps—as defined, e.g., in Ref. 17—which constitute the 15-parameter Lie group on Minkowski space. Liouville gave a proof of this theorem for three dimensions and Lie stated the theorem (*without proof*) for arbitrary dimension ≥ 3. A very explicit proof may be found in Beez[13] and the most palatable proof of this classical result seems to be the one given by Carathéodory.[18] The only diffeomorphisms among the similtudes and inversions on Minkowski space are the Lorentz transformations or dilatations by a constant. Hence we may summarize and obtain a very unified result for Minkowski space.

Corollary 6.5: Let M be the Minkowski space [with $\dim(M) \geq 3$]. For a map $h: M \to M$, the following are equivalent.

(1) h is a $\mathfrak{P}_i$-homeomorphism on M for one i, or a $\mathfrak{P}_i^*$-homeomorphism on M for one i, or a $\mathfrak{Z}_i$-homeomorphism on M for one i, or a $\mathfrak{Z}_i^*$-homeomorphism on M for one i, for $i \in \{0, 1, 2, \ldots, \infty, \omega, g\}$,

(2) h is a $\mathfrak{P}_i$-homeomorphism, $\mathfrak{P}_i^*$-homeomorphism, $\mathfrak{Z}_i$-homeomorphism and $\mathfrak{Z}_i^*$-homeomorphism on M for each $i = 0, 1, 2, \ldots, \infty, \omega, g$.

(3) h is a Lorentz transformation times a (linear) dilation with a constant > 0.

Remarks added in proof: It might be interesting to note, that the analog of Proposition 6.1 in the case of analytic Riemannian manifolds will be proved by Lelong-Ferrand[19] (Corollaire to Théorème A): Conformal C^2 maps are necessarily analytic. A "global proof" including Euclidean *and* hyperbolic metrics seems to be unknown.

[1]E. C. Zeeman, Topology 6, 161 (1967).
[2]R. Göbel, Commun. Math. Phys. 43, 289 (1976).
[3]S. W. Hawking, A. R. King, and P. J. McCarthy, *"A New Topology for Curved Space—Time which Incorporates the Causal, Differential, and Conformal Structures,"* to appear in J. Math. Phys.
[4]R. Penrose, *Techniques of Differential Topology in Relativity* (Soc. Ind. Appl. Math., Philadelphia, 1972).
[5]A. Einstein, Universidad Nacional de Tucuman Revista A 2, 11 (1941).
[6]J. Winicour, Springer Lecture Notes in Physics 14, 145 (1972).
[7]S. Nanda, J. Math. Phys. 12, 394 (1971); 13, 12 (1972).
[8]R. Geroch, Math. Rev. 43, 808 (1972).
[9]S. W. Hawking and G. F. R. Ellis, *The Large Scale Structure of Space—Time* (Cambridge U. P., Cambridge, 1973).
[10]J. Liouville, J. Math. Pures Appl. 12, 265 (1847).
[11]N. J. Hicks, *Notes on Differential Geometry* (Van Nostrand, London, 1965).
[12]R. Narasimhan, *Analysis on Real and Complex Manifolds* (North-Holland, Amsterdam, 1968).
[13]R. Beez, Math. Phys. (Leipz) 20, 253 (1875).
[14]R. Göbel, *Die volle kausale Gruppe der Raum-Zeit*, Physikal Teil II der Habilitationsschrift, Würzburg, 1973.
[15]S. W. Hawking, *Singularities and the Geometry of Space Time,"* Adams Prize Essay, Cambridge (1966).
[16]S. Lie, Math. Ann. 5, 145 (1872), cf. p. 186.
[17]S. Ferrara, R. Gatto, and A. F. Grillo, Springer Tracts Mod. Phys. 67, xxx (1973).
[18]C. Carathéodory, Sitzungsber. der Preussischen Akad. d. Wiss., phys.-math. Klasse Sec. 25, 12 (1924).
[19]J. Lelong-Ferrand, "Interpretations geometriques de la courbure scalaire et regularite des homoeomorphismes conformes," to appear in the jubilee book, dedicated to A. Lichnerowicz (Reidel, Brussels, to be published).

ERRATA

Erratum: New Jacobian θ functions and the evaluation of lattice sums [J. Math. Phys. 16, 2189 (1975)]

I. J. Zucker

Department of Physics, University of Surrey, Guilford, Surrey, England

(Received 2 January 1976)

On p. 2190, bottom left-hand column: The line reading "$= 2^s[L_{8a}(s) + L_{8b}(s)]$" should read "$= 2^s[L_{8a}(s) \times L_{8b}(s)]$."

On p. 2190, top right-hand column: The line reading "$= [\pi + 2\ln(1+\sqrt{2})]/\sqrt{2}$" should read "$= \pi\sqrt{2}\ln(1+\sqrt{2})$."

Corollary 6.5: Let M be the Minkowski space [with $\dim(M) \geq 3$]. For a map $h: M \to M$, the following are equivalent.

(1) h is a $\mathfrak{P}_i$-homeomorphism on M for one i, or a $\mathfrak{P}_i^*$-homeomorphism on M for one i, or a $\mathfrak{Z}_i$-homeomorphism on M for one i, or a $\mathfrak{Z}_i^*$-homeomorphism on M for one i, for $i \in \{0, 1, 2, \ldots, \infty, \omega, g\}$,

(2) h is a $\mathfrak{P}_i$-homeomorphism, $\mathfrak{P}_i^*$-homeomorphism, $\mathfrak{Z}_i$-homeomorphism and $\mathfrak{Z}_i^*$-homeomorphism on M for each $i = 0, 1, 2, \ldots, \infty, \omega, g$.

(3) h is a Lorentz transformation times a (linear) dilation with a constant > 0.

Remarks added in proof: It might be interesting to note, that the analog of Proposition 6.1 in the case of analytic Riemannian manifolds will be proved by Lelong-Ferrand[19] (Corollaire to Théorème A): Conformal C^2 maps are necessarily analytic. A "global proof" including Euclidean *and* hyperbolic metrics seems to be unknown.

[1]E. C. Zeeman, Topology 6, 161 (1967).
[2]R. Göbel, Commun. Math. Phys. 43, 289 (1976).
[3]S. W. Hawking, A. R. King, and P. J. McCarthy, *"A New Topology for Curved Space—Time which Incorporates the Causal, Differential, and Conformal Structures,"* to appear in J. Math. Phys.
[4]R. Penrose, *Techniques of Differential Topology in Relativity* (Soc. Ind. Appl. Math., Philadelphia, 1972).
[5]A. Einstein, Universidad Nacional de Tucuman Revista A 2, 11 (1941).
[6]J. Winicour, Springer Lecture Notes in Physics 14, 145 (1972).
[7]S. Nanda, J. Math. Phys. 12, 394 (1971); 13, 12 (1972).
[8]R. Geroch, Math. Rev. 43, 808 (1972).
[9]S. W. Hawking and G. F. R. Ellis, *The Large Scale Structure of Space—Time* (Cambridge U. P., Cambridge, 1973).
[10]J. Liouville, J. Math. Pures Appl. 12, 265 (1847).
[11]N. J. Hicks, *Notes on Differential Geometry* (Van Nostrand, London, 1965).
[12]R. Narasimhan, *Analysis on Real and Complex Manifolds* (North-Holland, Amsterdam, 1968).
[13]R. Beez, Math. Phys. (Leipz) 20, 253 (1875).
[14]R. Göbel, *Die volle kausale Gruppe der Raum-Zeit*, Physikal Teil II der Habilitationsschrift, Würzburg, 1973.
[15]S. W. Hawking, *Singularities and the Geometry of Space Time,"* Adams Prize Essay, Cambridge (1966).
[16]S. Lie, Math. Ann. 5, 145 (1872), cf. p. 186.
[17]S. Ferrara, R. Gatto, and A. F. Grillo, Springer Tracts Mod. Phys. 67, xxx (1973).
[18]C. Carathéodory, Sitzungsber. der Preussischen Akad. d. Wiss., phys.-math. Klasse Sec. 25, 12 (1924).
[19]J. Lelong-Ferrand, "Interpretations geometriques de la courbure scalaire et regularite des homoeomorphismes conformes," to appear in the jubilee book, dedicated to A. Lichnerowicz (Reidel, Brussels, to be published).

ERRATA

Erratum: New Jacobian θ functions and the evaluation of lattice sums [J. Math. Phys. 16, 2189 (1975)]

I. J. Zucker

Department of Physics, University of Surrey, Guilford, Surrey, England

(Received 2 January 1976)

On p. 2190, bottom left-hand column: The line reading "$= 2^s[L_{8a}(s) + L_{8b}(s)]$" should read "$= 2^s[L_{8a}(s) \times L_{8b}(s)]$."

On p. 2190, top right-hand column: The line reading "$= [\pi + 2\ln(1+\sqrt{2})]/\sqrt{2}$" should read "$= \pi\sqrt{2}\ln(1+\sqrt{2})$."